In [1]:
```
library(tidyverse)
```

```
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Registered S3 method overwritten by 'rvest':
  method            from
  read_xml.response xml2
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.1.1       ✔ purrr   0.3.2
✔ tibble  2.1.1       ✔ dplyr   0.8.0.1
✔ tidyr   0.8.3       ✔ stringr 1.4.0
✔ readr   1.3.1       ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

In [4]:
```
# what is the working directory?
getwd()
```

'/Users/mruddy/Documents/BSDS100'

In [6]:
```
# read a csv file inside of the data folder
# (which should lie in the working directory)
avocado_data <- read_csv("data/avocado.csv")
```

```
Warning message:
"Missing column names filled in: 'X1' [1]"Parsed with column specification:
cols(
  X1 = col_double(),
  Date = col_date(format = ""),
  AveragePrice = col_double(),
  `Total Volume` = col_double(),
  `4046` = col_double(),
  `4225` = col_double(),
  `4770` = col_double(),
  `Total Bags` = col_double(),
  `Small Bags` = col_double(),
  `Large Bags` = col_double(),
  `XLarge Bags` = col_double(),
  type = col_character(),
  year = col_double(),
  region = col_character()
)
```

In [8]:
```
print(avocado_data)
```

```
# A tibble: 18,249 x 14
      X1 Date       AveragePrice `Total Volume` `4046` `4225` `4770`
   <dbl> <date>            <dbl>          <dbl>  <dbl>  <dbl>  <dbl>
 1     0 2015-12-27         1.33         64237.  1037. 5.45e4   48.2
 2     1 2015-12-20         1.35         54877.   674. 4.46e4   58.3
 3     2 2015-12-13         0.93        118220.   795. 1.09e5  130.
 4     3 2015-12-06         1.08         78992.  1132  7.20e4   72.6
 5     4 2015-11-29         1.28         51040.   941. 4.38e4   75.8
 6     5 2015-11-22         1.26         55980.  1184. 4.81e4   43.6
 7     6 2015-11-15         0.99         83454.  1369. 7.37e4   93.3
 8     7 2015-11-08         0.98        109428.   704. 1.02e5   80
 9     8 2015-11-01         1.02         99811.  1022. 8.73e4   85.3
10     9 2015-10-25         1.07         74339.   842. 6.48e4  113
# … with 18,239 more rows, and 7 more variables: `Total Bags` <dbl>, `Small
#   Bags` <dbl>, `Large Bags` <dbl>, `XLarge Bags` <dbl>, type <chr>,
#   year <dbl>, region <chr>
```

In [10]:
```
expensive_avocados <- avocado_data %>%
  filter(AveragePrice > 1.50) %>%
  print()
```

```
# A tibble: 6,912 x 14
      X1 Date       AveragePrice `Total Volume` `4046` `4225` `4770`
   <dbl> <date>            <dbl>          <dbl>  <dbl>  <dbl>  <dbl>
 1    14 2015-09-20         1.54         60624. 1.45e3 3.98e4 8.86e1
 2    15 2015-09-13         1.59         73043. 1.57e3 4.85e4 1.34e2
 3    16 2015-09-06         1.56         76140. 1.47e3 4.94e4 1.73e2
 4    49 2015-01-18         1.52        107040. 1.63e3 6.19e4 1.33e2
 5    50 2015-01-11         1.54        106221. 1.64e3 5.10e4 7.96e1
 6    50 2015-01-11         1.54         54644. 1.49e3 3.38e4 1.33e3
 7    51 2015-01-04         1.52         54957. 3.01e3 3.55e4 1.56e3
 8     9 2015-10-25         1.55        561342. 1.26e5 3.34e5 3.14e4
 9    10 2015-10-18         1.52        586074. 1.47e5 3.33e5 3.10e4
10    11 2015-10-11         1.58        564457. 1.06e5 3.52e5 3.35e4
# … with 6,902 more rows, and 7 more variables: `Total Bags` <dbl>, `Small
#   Bags` <dbl>, `Large Bags` <dbl>, `XLarge Bags` <dbl>, type <chr>,
#   year <dbl>, region <chr>
```

In [11]:
```
# write a csv file
write_csv(expensive_avocados, "data/expensive_avocados.csv")
```

In [22]:
```
getwd()
```

'/Users/mruddy/Documents/BSDS100'

The following dataset can be found here: https://www.kaggle.com/sakshigoyal7/credit-card-customers

In [19]:
```
creditcard_data <- read_csv("data/BankChurners.csv")
```

```
Parsed with column specification:
cols(
  .default = col_double(),
  Attrition_Flag = col_character(),
  Gender = col_character(),
  Education_Level = col_character(),
  Marital_Status = col_character(),
  Income_Category = col_character(),
  Card_Category = col_character()
)
See spec(...) for full column specifications.
```

In [28]:
```
print(creditcard_data, width = Inf, n = 5)
```

```
# A tibble: 10,127 x 23
  CLIENTNUM Attrition_Flag    Customer_Age Gender Dependent_count
      <dbl> <chr>                    <dbl> <chr>            <dbl>
1 768805383 Existing Customer           45 M                    3
2 818770008 Existing Customer           49 F                    5
3 713982108 Existing Customer           51 M                    3
4 769911858 Existing Customer           40 F                    4
5 709106358 Existing Customer           40 M                    3
  Education_Level Marital_Status Income_Category Card_Category Months_on_book
  <chr>           <chr>          <chr>           <chr>                  <dbl>
1 High School     Married        $60K - $80K     Blue                      39
2 Graduate        Single         Less than $40K  Blue                      44
3 Graduate        Married        $80K - $120K    Blue                      36
4 High School     Unknown        Less than $40K  Blue                      34
5 Uneducated      Married        $60K - $80K     Blue                      21
  Total_Relationship_Count Months_Inactive_12_mon Contacts_Count_12_mon
                     <dbl>                  <dbl>                 <dbl>
1                        5                      1                     3
2                        6                      1                     2
3                        4                      1                     0
4                        3                      4                     1
5                        5                      1                     0
  Credit_Limit Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1
         <dbl>               <dbl>           <dbl>                <dbl>
1        12691                 777           11914                 1.34
2         8256                 864            7392                 1.54
3         3418                   0            3418                 2.59
4         3313                2517             796                 1.40
5         4716                   0            4716                 2.17
  Total_Trans_Amt Total_Trans_Ct Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
            <dbl>          <dbl>               <dbl>                 <dbl>
1            1144             42                1.62                 0.061
2            1291             33                3.71                 0.105
3            1887             20                2.33                 0
4            1171             20                2.33                 0.76
5             816             28                2.5                  0
  Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dep…
                                                                           <dbl>
1                                                                      0.0000934
2                                                                      0.0000569
3                                                                      0.0000211
4                                                                      0.000134
5                                                                      0.0000217
  Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dep…
                                                                           <dbl>
1                                                                          1.000
2                                                                          1.000
3                                                                          1.000
4                                                                          1.000
5                                                                          1.000
# … with 1.012e+04 more rows
```

In [30]:
```
creditcard_data %>%
  group_by(Income_Category) %>%
  summarize(mean = mean(Customer_Age, na.rm = TRUE))
```

Income_Category	mean
$120K +	47.60385
40*K*−60K	46.08715
60*K*−80K	45.96862
80*K*−120K	46.42801
Less than $40K	46.29795
Unknown	46.27428

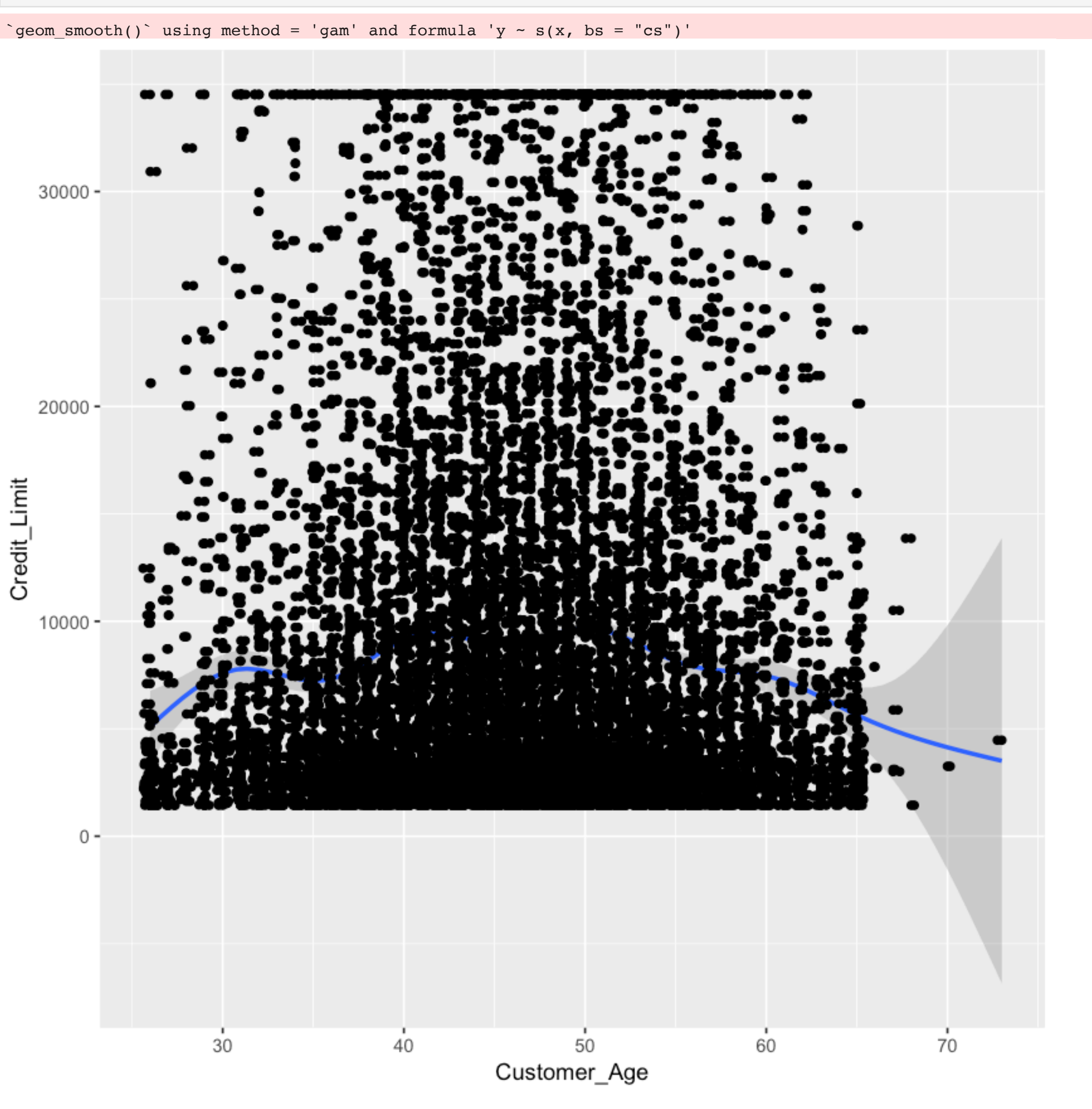
In [33]:
```
creditcard_data %>%
  ggplot(aes(x = Customer_Age)) +
  geom_freqpoly(aes(color = Income_Category))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Customer_Age, Credit_Limit

In [37]:
```
creditcard_data %>%
  ggplot(aes(x = Customer_Age, y = Credit_Limit)) +
  geom_smooth() +
  geom_point() +
  geom_jitter()
```

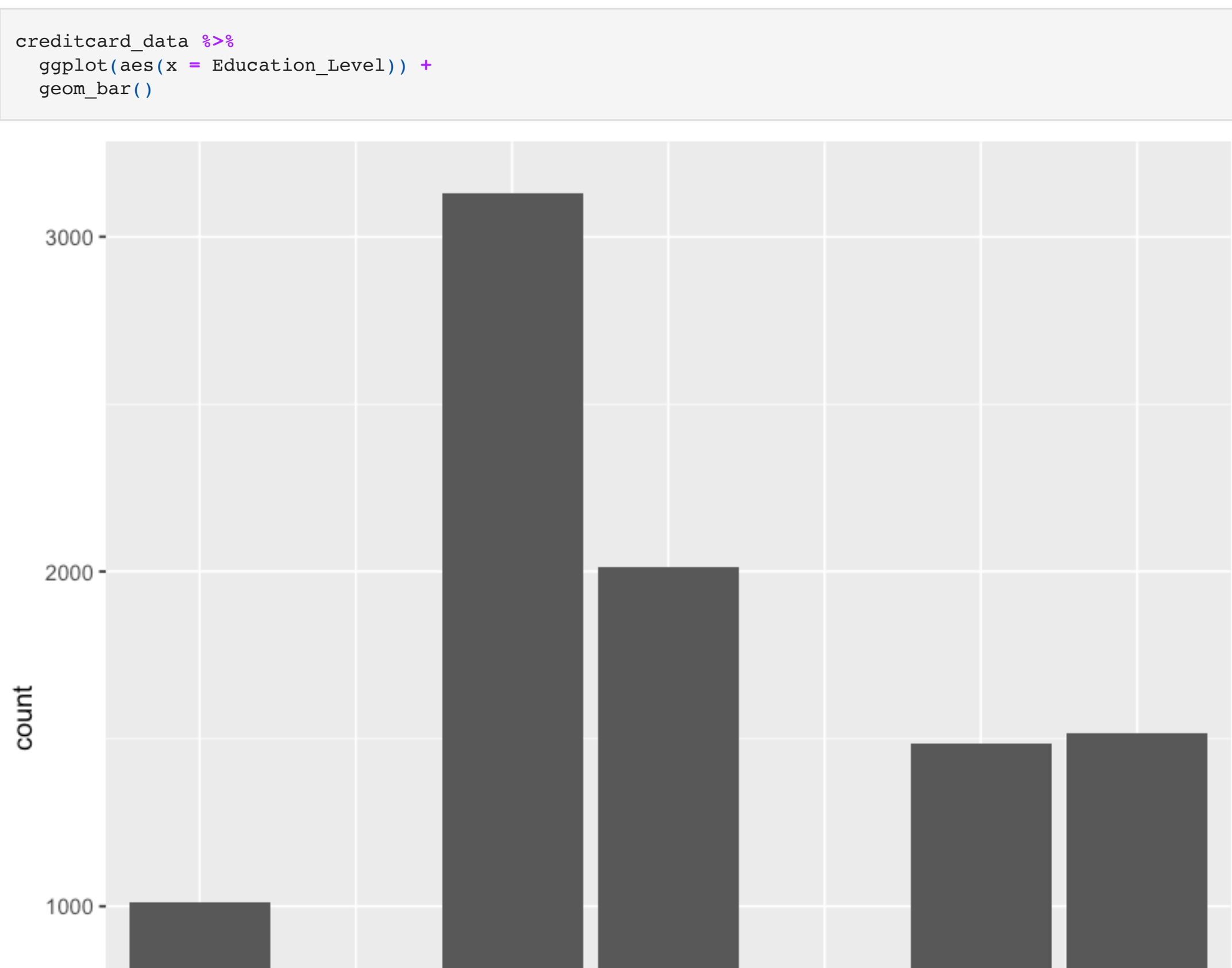
```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



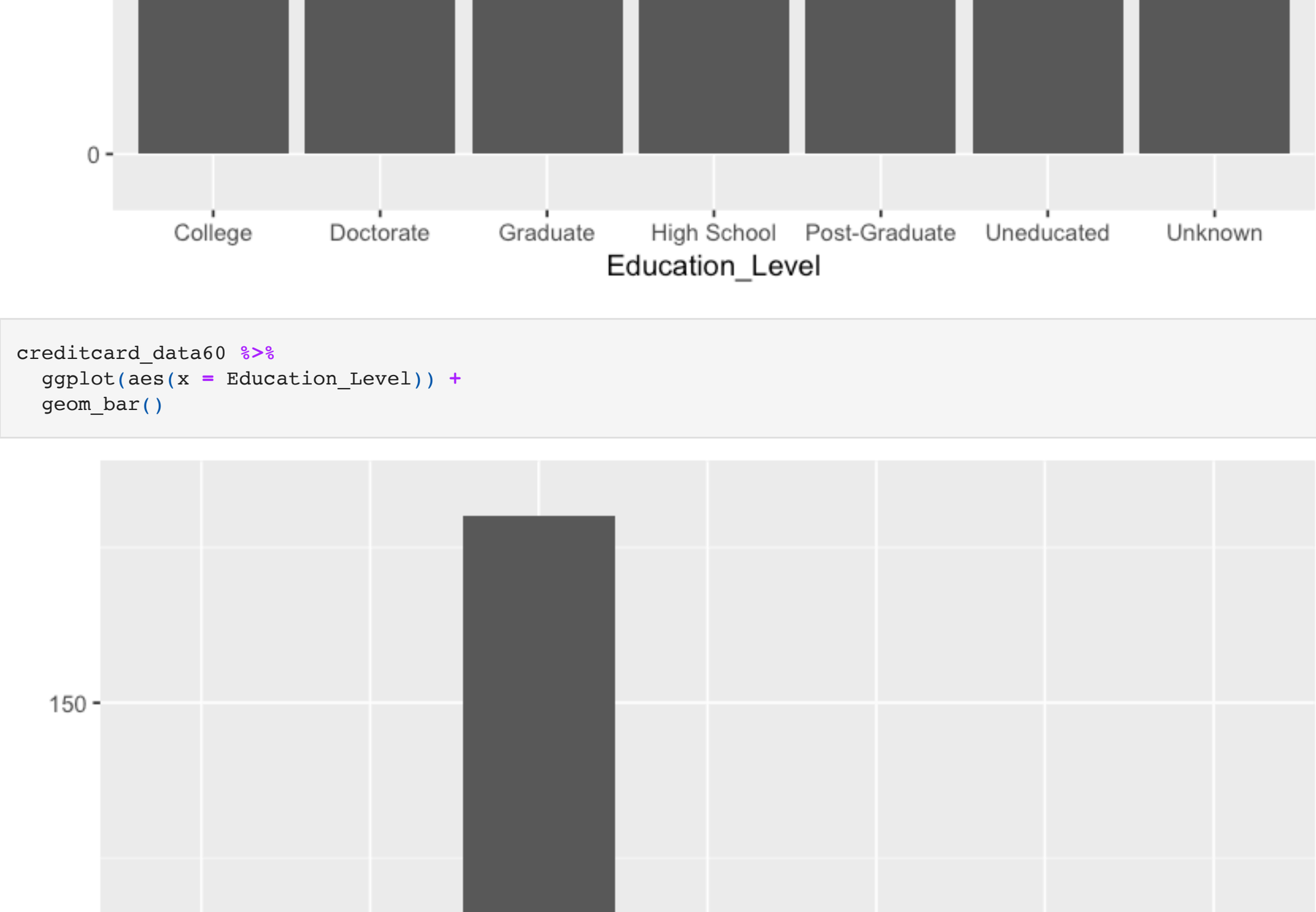
In [38]:
```
creditcard_data60 <- creditcard_data %>%
  filter(Customer_Age >= 60) %>%
  print()
```

```
# A tibble: 532 x 23
   CLIENTNUM Attrition_Flag Customer_Age Gender Dependent_count Education_Level
       <dbl> <chr>                 <dbl> <chr>            <dbl> <chr>
 1 710821833 Existing Cust…           65 M                    1 Unknown
 2 806160108 Existing Cust…           61 M                    1 High School
 3 708508758 Attrited Cust…           62 F                    0 Graduate
 4 804424383 Existing Cust…           63 M                    1 Unknown
 5 708300483 Attrited Cust…           66 F                    0 Doctorate
 6 711525033 Existing Cust…           66 F                    0 High School
 7 719712633 Existing Cust…           64 M                    1 Graduate
 8 808284783 Existing Cust…           62 F                    1 Unknown
 9 711112683 Existing Cust…           63 F                    1 College
10 712720158 Existing Cust…           68 M                    1 Graduate
# … with 522 more rows, and 17 more variables: Marital_Status <chr>,
#   Income_Category <chr>, Card_Category <chr>, Months_on_book <dbl>,
#   Total_Relationship_Count <dbl>, Months_Inactive_12_mon <dbl>,
#   Contacts_Count_12_mon <dbl>, Credit_Limit <dbl>, Total_Revolving_Bal <dbl>,
#   Avg_Open_To_Buy <dbl>, Total_Amt_Chng_Q4_Q1 <dbl>, Total_Trans_Amt <dbl>,
#   Total_Trans_Ct <dbl>, Total_Ct_Chng_Q4_Q1 <dbl>,
#   Avg_Utilization_Ratio <dbl>,
#   Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1 <dbl>,
#   Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2 <dbl>
```

In [39]:
```
creditcard_data %>%
  ggplot(aes(x = Education_Level)) +
  geom_bar()
```



In [40]:
```
creditcard_data60 %>%
  ggplot(aes(x = Education_Level)) +
  geom_bar()
```



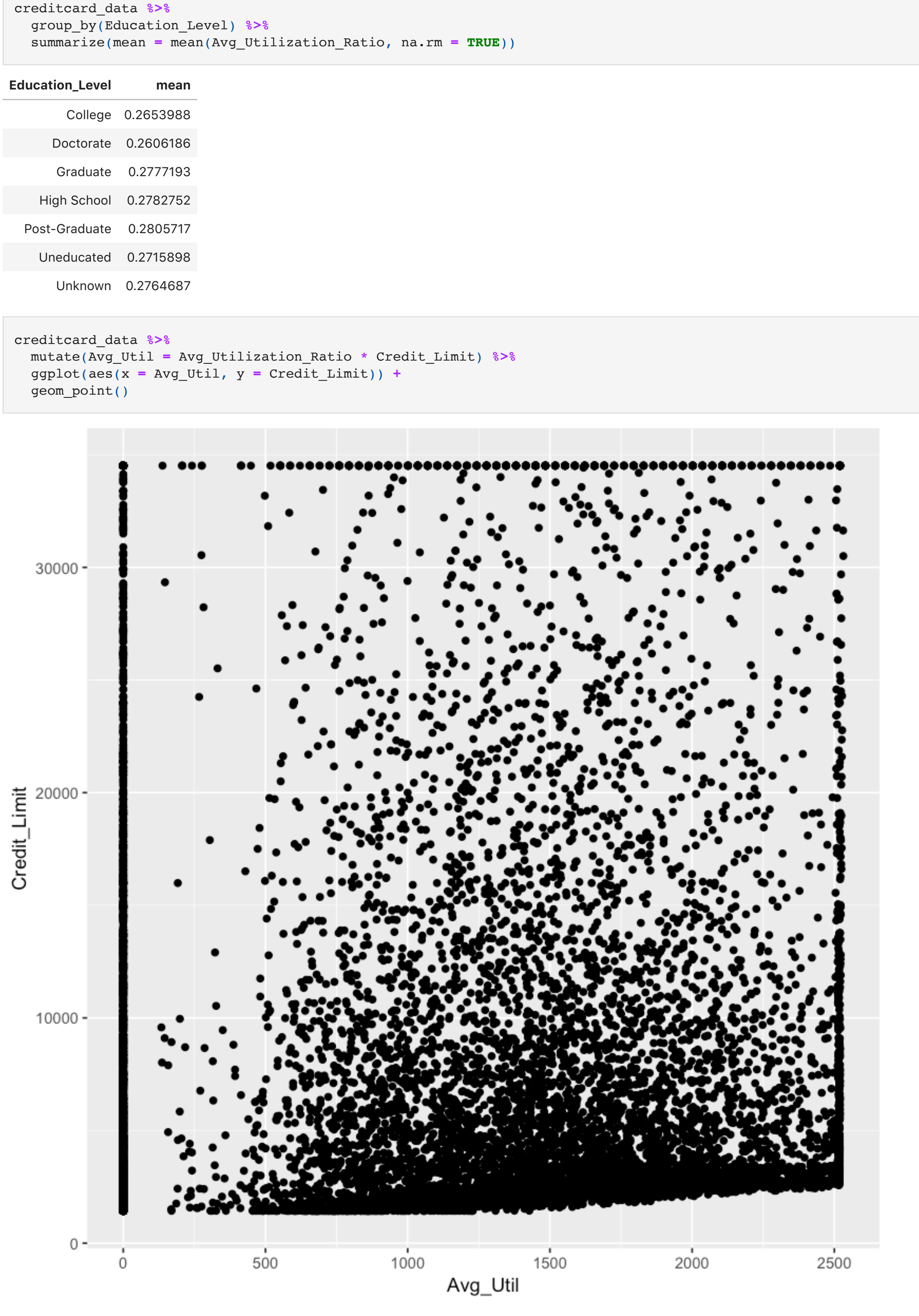
In [41]:
```
creditcard_data %>%
  ggplot(aes(x = Avg_Utilization_Ratio)) +
  geom_freqpoly(aes(color = Education_Level))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

In [42]:
```
creditcard_data %>%
  group_by(Education_Level) %>%
  summarize(mean = mean(Avg_Utilization_Ratio, na.rm = TRUE))
```

Education_Level	mean
College	0.2653988
Doctorate	0.2606186
Graduate	0.2777193
High School	0.2782752
Post-Graduate	0.2805717
Uneducated	0.2715898
Unknown	0.2764687

In [44]:
```
creditcard_data %>%
  mutate(Avg_Util = Avg_Utilization_Ratio * Credit_Limit) %>%
  ggplot(aes(x = Avg_Util, y = Credit_Limit)) +
  geom_point()
```



In [ ]: