02 - Visualisation de données - Partie 2

PRO1036 - Analyse de données scientifiques en R

Tim Bollé

September 9, 2024

Visualisation des données

Terminologie

Analyse :

- `Univariée` - distribution d'une unique variable
- `Bivariée` - Relation entre deux variables
- `Multivariée` - Relation entre plusieurs variables, souvent en se concentrant sur la relation entre deux, tout en les conditionnant selon les autres.

Terminologie

Types de variables :

		Opérations possibles
Qualitative	Nominale	= ≠
	Ordinale	= ≠ < >
Quantitative	Intervalle	= ≠ < > + -
	Ratio	= ≠ < > + - · /

Données

Lending Club

Plateforme pour faire des prêts entre particuliers.

Le jeu de données contient les prêts effectués.

```
1 library(openintro)
2 glimpse(loans_full_schema)
```

```
Rows: 10,000
Columns: 55
$ emp_title                    <chr> "global config engineer ", "warehouse…
$ emp_length                   <dbl> 3, 10, 3, 1, 10, NA, 10, 10, 10, 3, 1…
$ state                        <fct> NJ, HI, WI, PA, CA, KY, MI, AZ, NV, I…
$ homeownership                <fct> MORTGAGE, RENT, RENT, RENT, RENT, OWN…
$ annual_income                <dbl> 90000, 40000, 40000, 30000, 35000, 34…
$ verified_income              <fct> Verified, Not Verified, Source Verifi…
$ debt_to_income               <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.4…
$ annual_income_joint          <dbl> NA, NA, NA, NA, 57000, NA, 155000, NA…
$ verification_income_joint    <fct> , , , , Verified, , Not Verified, , ,…
$ debt_to_income_joint         <dbl> NA, NA, NA, NA, 37.66, NA, 13.12, NA,
```

Sélection de variables

```
1 loans <- loans_full_schema %>%
2   select(loan_amount, interest_rate, term, grade,
3          state, annual_income, homeownership, debt_to_income)
4 glimpse(loans)
```

```
Rows: 10,000
Columns: 8
$ loan_amount    <int> 28000, 5000, 2000, 21600, 23000, 5000, 24000, 20000, 20…
$ interest_rate  <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.72, 13.59, 11.99, 1…
$ term           <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36, 36, 60, 60, 36, 60,…
$ grade          <fct> C, C, D, A, C, A, C, B, C, A, C, B, C, B, D, D, D, F, E…
$ state          <fct> NJ, HI, WI, PA, CA, KY, MI, AZ, NV, IL, IL, FL, SC, CO,…
$ annual_income  <dbl> 90000, 40000, 40000, 30000, 35000, 34000, 35000, 110000…
$ homeownership  <fct> MORTGAGE, RENT, RENT, RENT, RENT, OWN, MORTGAGE, MORTGA…
$ debt_to_income <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.46, 23.66, 16.19, 3…
```

Variables sélectionnées

Variable	Description
`loan_amount`	Montant du prêt reçu en US Dollards
`interest_rate`	Intérêt sur le prêt, en pourcentage annuel
`term`	Durée du prêt en mois
`grade`	Note du prêt, de A à G, qui représente la qualité du prêt et les changes qu'ils soit remboursé
`state`	État américain dans lequel le prêt a été accordé
`annual_income`	Revenu annuel dudébiteur, en US Dollards
`homeownership`	Indique si la personne est propriétaire, est propriétaire avec un emprunt ou bien loue sa résidence
`debt_to_income`	Ratio Dette/Revenu

Types des variables

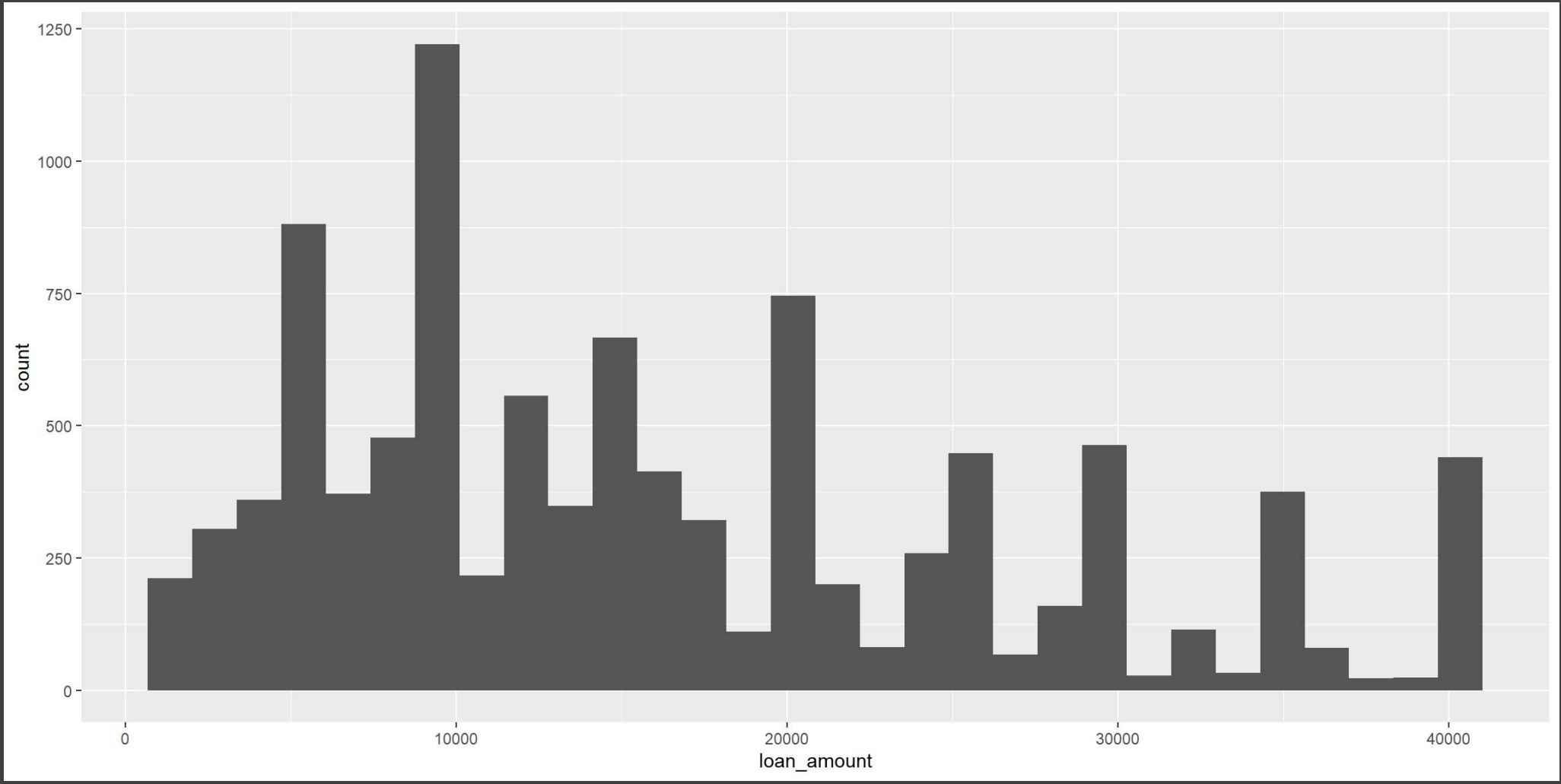
variable	type
`loan_amount`	Quantitatif, Ratio
`interest_rate`	Quantitatif, Ratio
`term`	Quantitatif, Ratio
`grade`	Qualitatif, Ordinal
`state`	Qualitatif, Nominal
`annual_income`	Quantitatif, Ratio
`homeownership`	Qualitatif, Nominal
`debt_to_income`	Quantitatif, Ratio

Données quantitatives

Histogramme

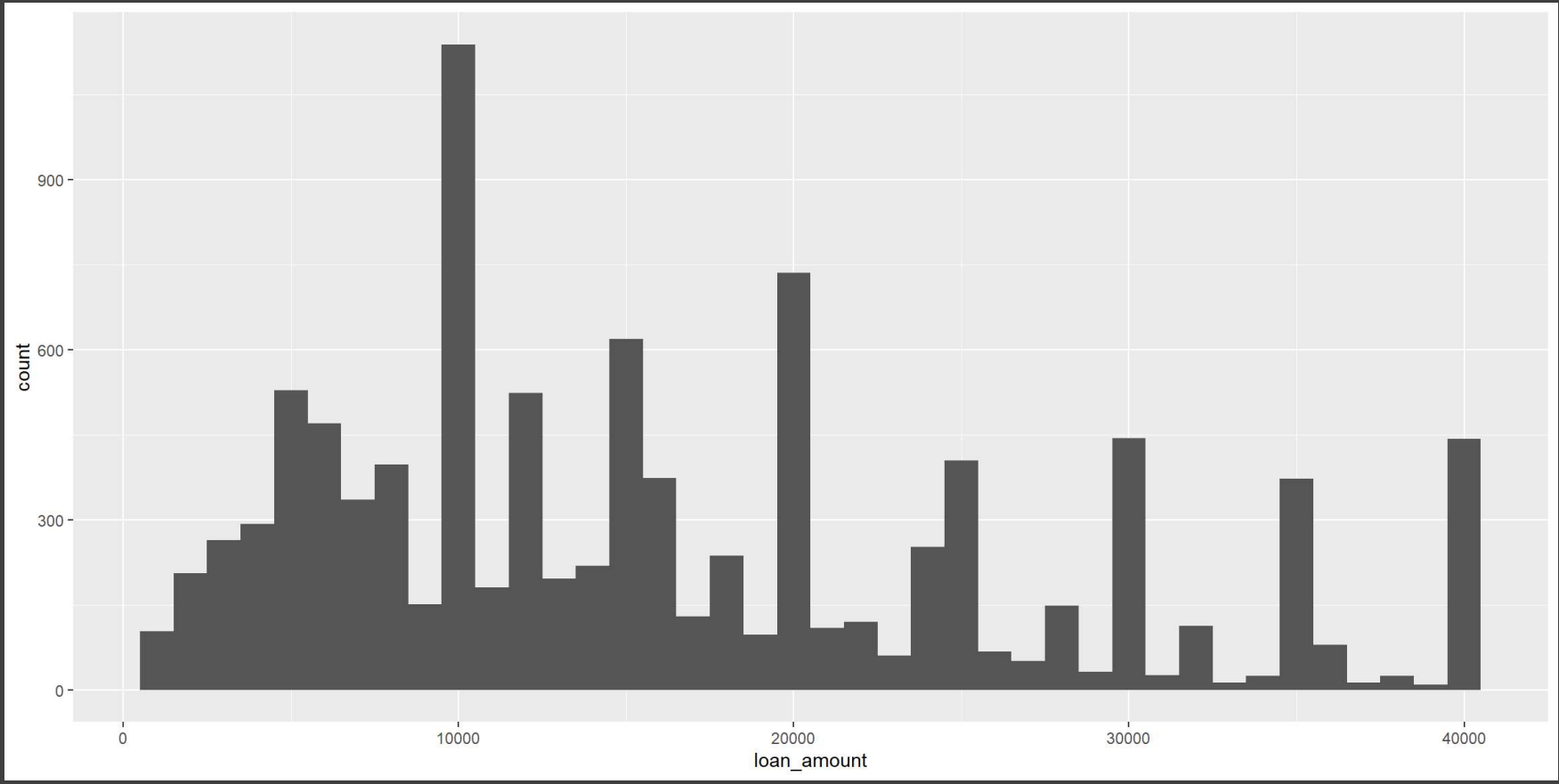
Histogramme

```
1 ggplot(loans, aes(x = loan_amount)) +
2   geom_histogram()
```



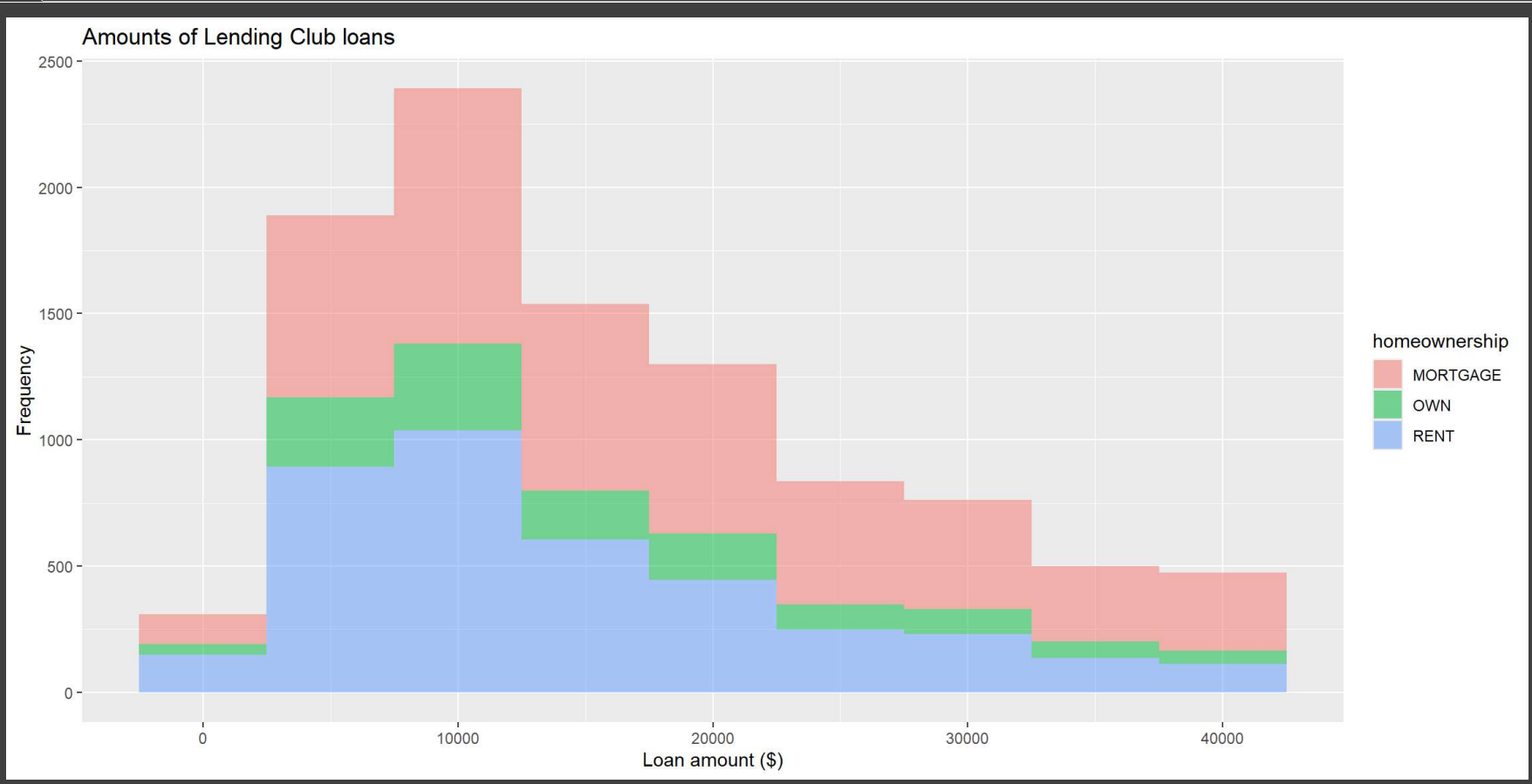
binwidth = 1000 | binwidth = 5000 | binwidth = 20000

```
1 ggplot(loans, aes(x = loan_amount)) +
2   geom_histogram(binwidth = 1000)
```



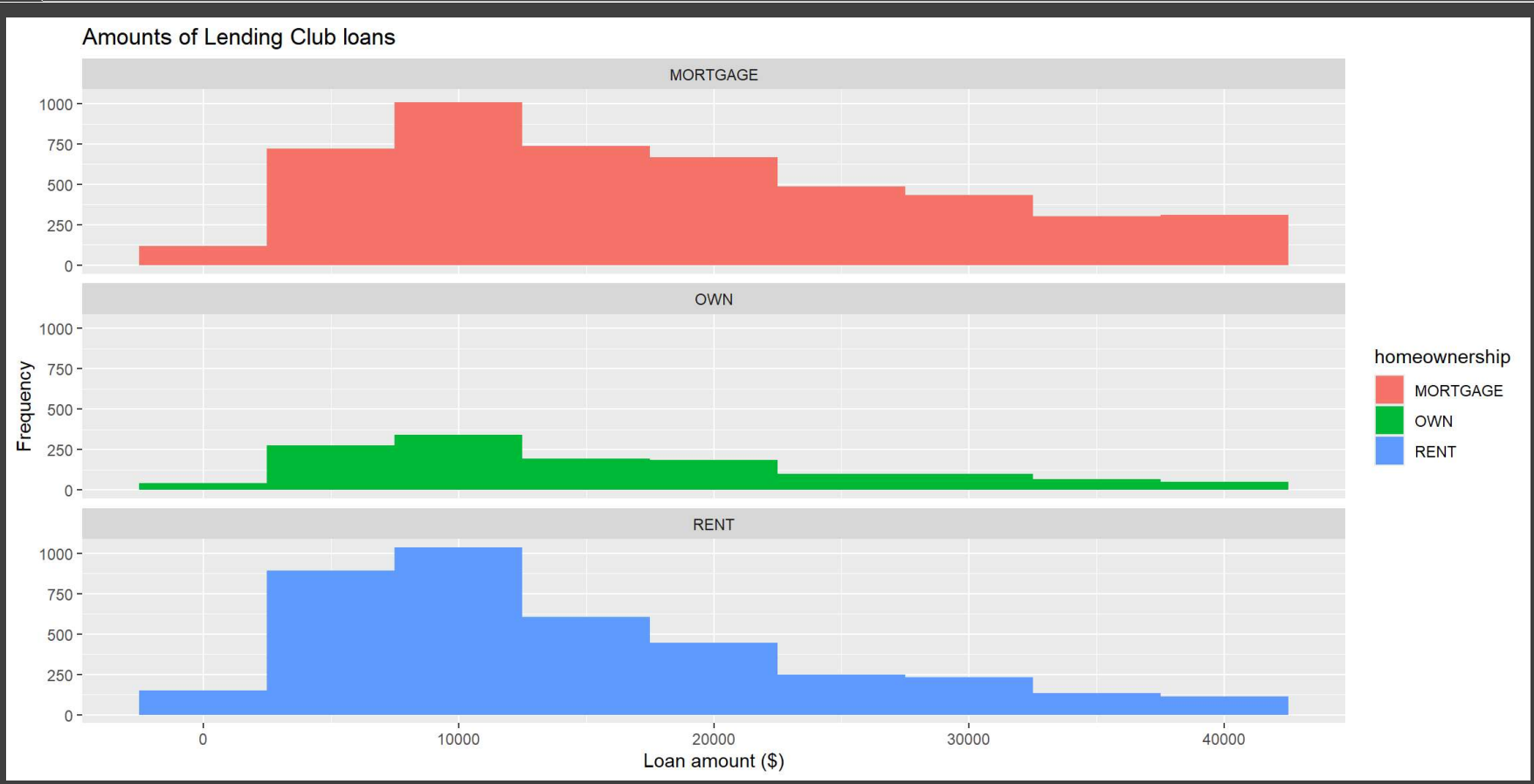
Combinaison avec des variables qualitatives

Plot | Code



Avec des facettes

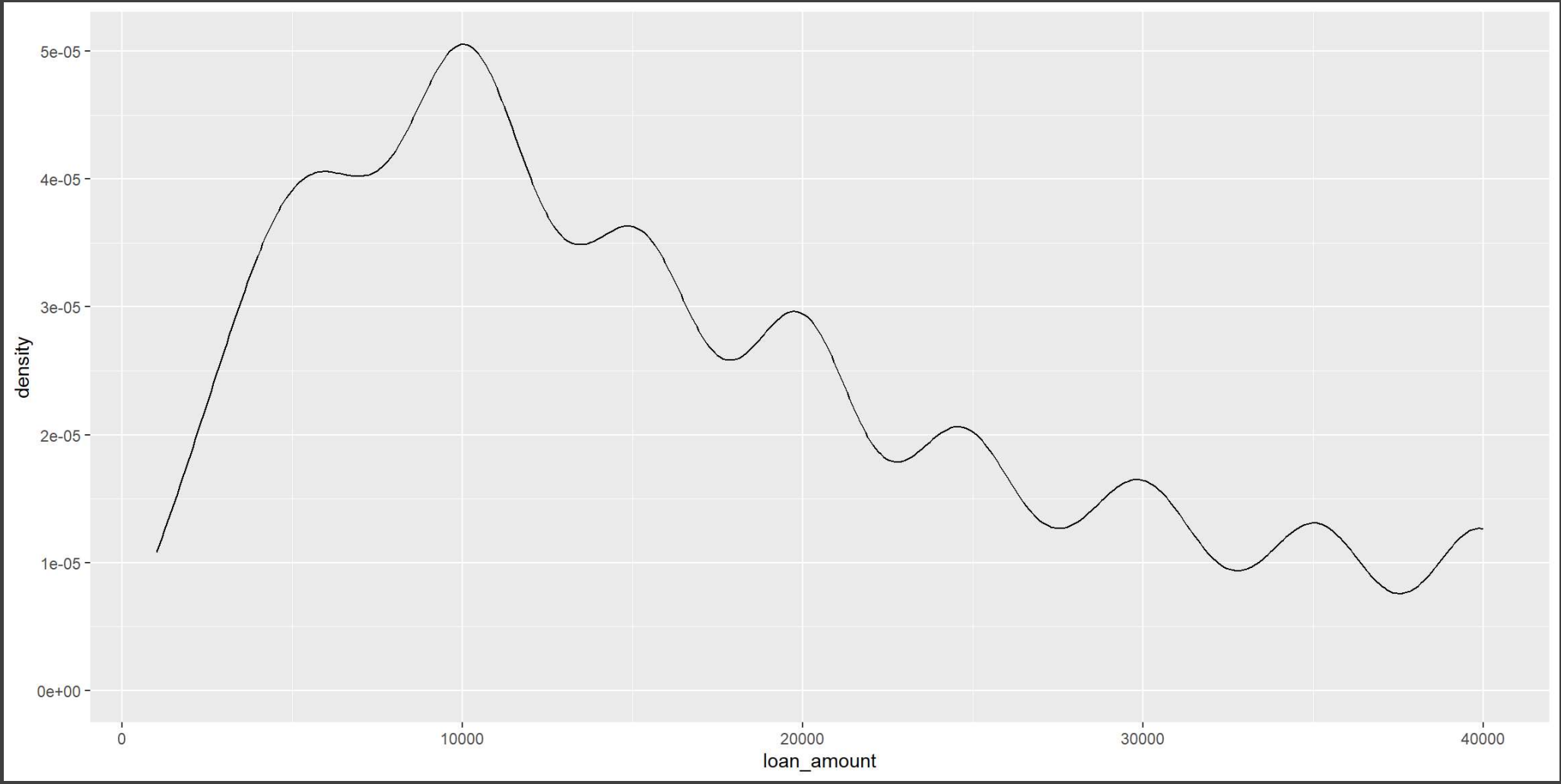
Plot | Code



Graphe de densité

Density plot

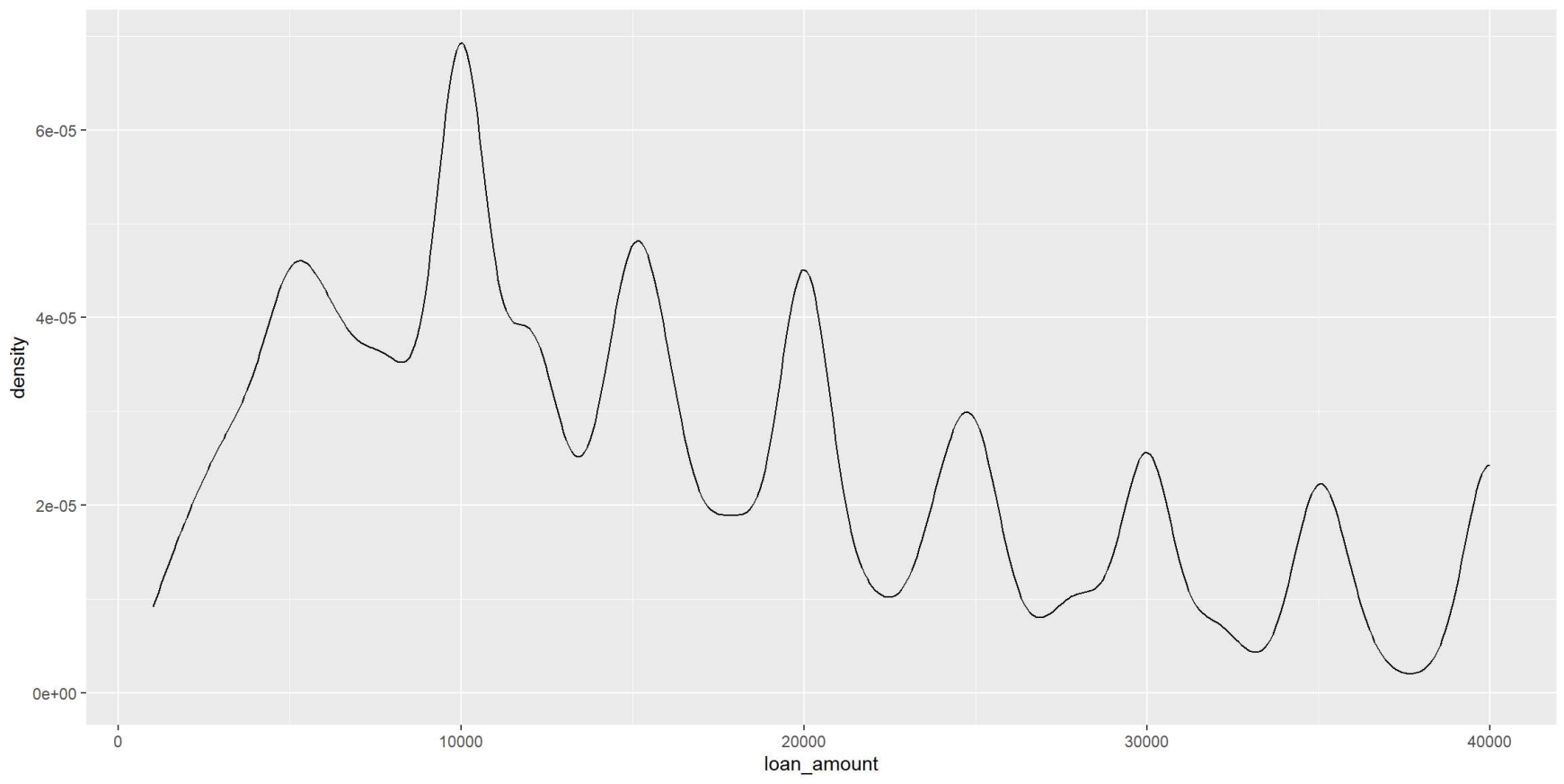
```
1  ggplot(loans, aes(x = loan_amount)) +
2    geom_density()
```



Ajustement de la précision

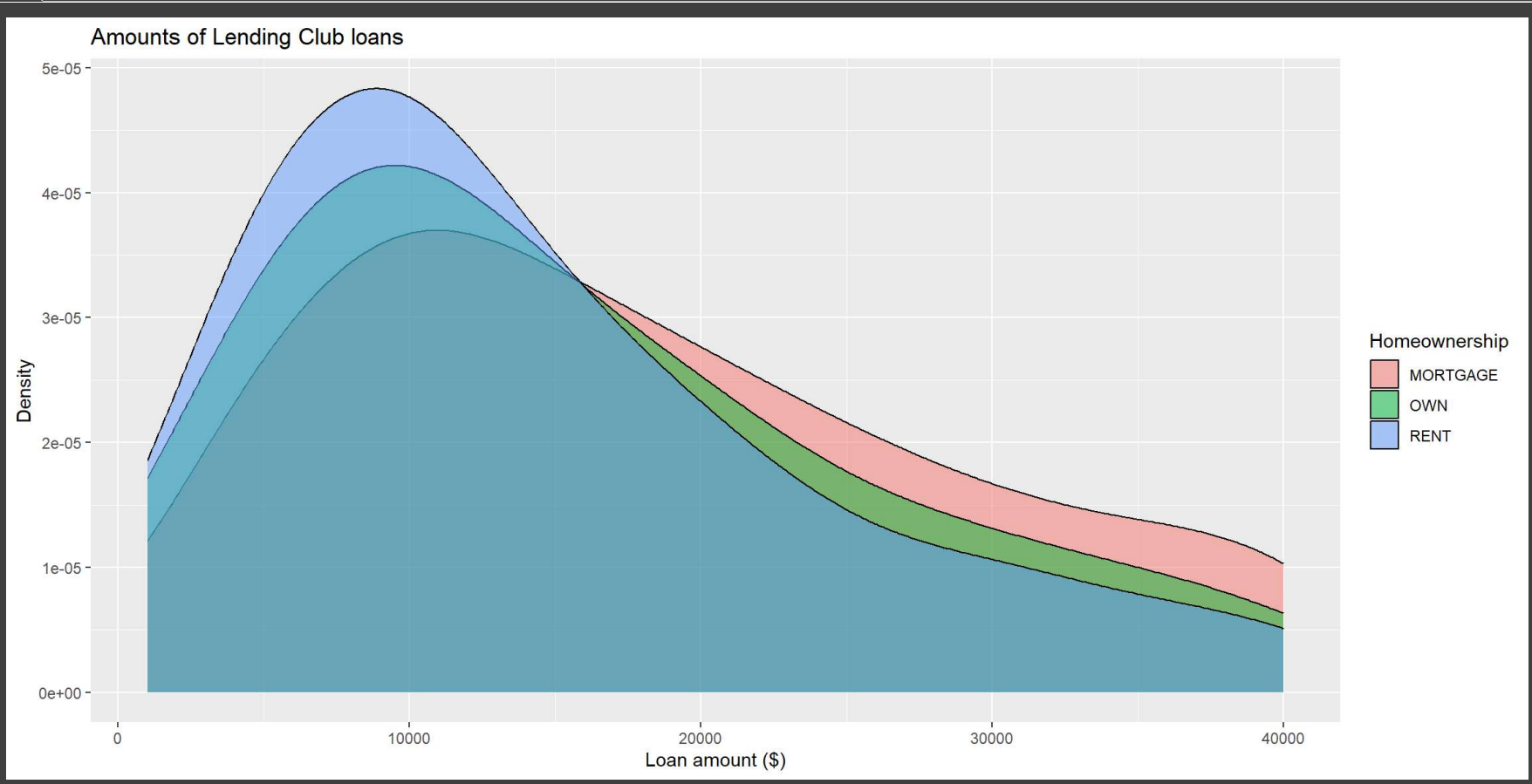
adjust = 0.5 | adjust = 1 | adjust = 2

```
1 ggplot(loans, aes(x = loan_amount)) +
2   geom_density(adjust = 0.5)
```



Combinaison avec des variables qualitatives

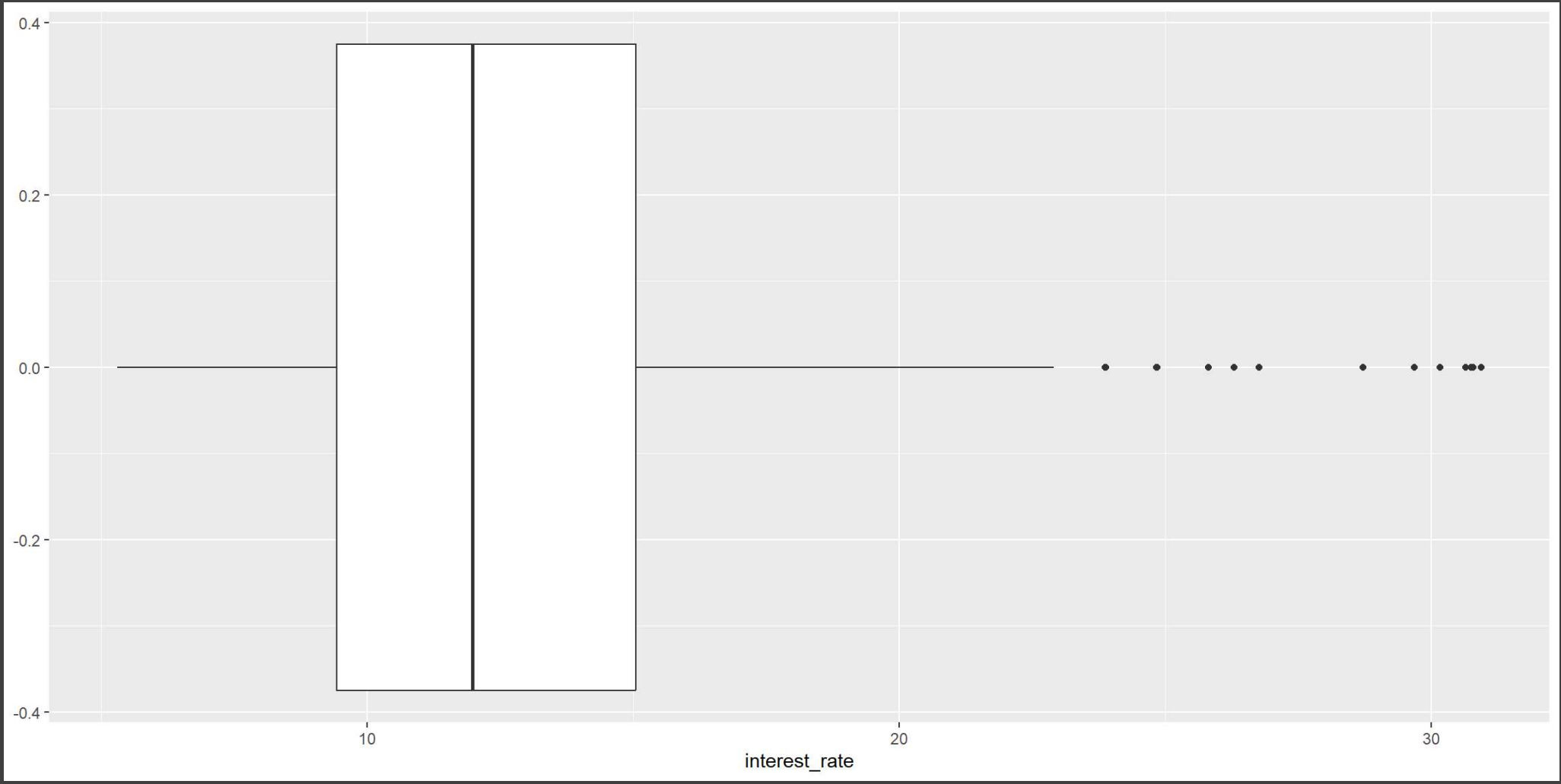
Plot | Code



Box plot

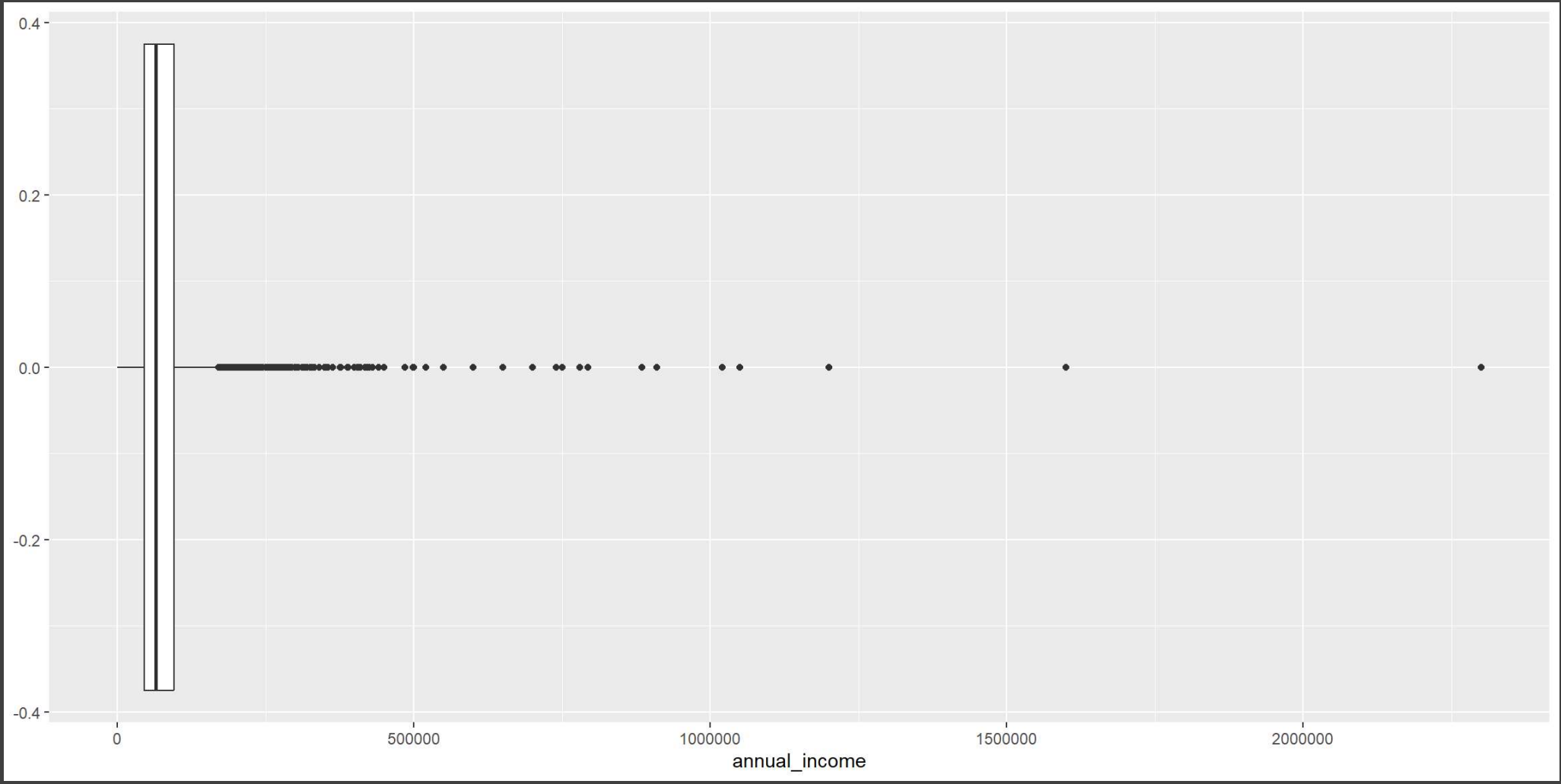
Boîte à moustache

```
1 ggplot(loans, aes(x = interest_rate)) +
2   geom_boxplot()
```



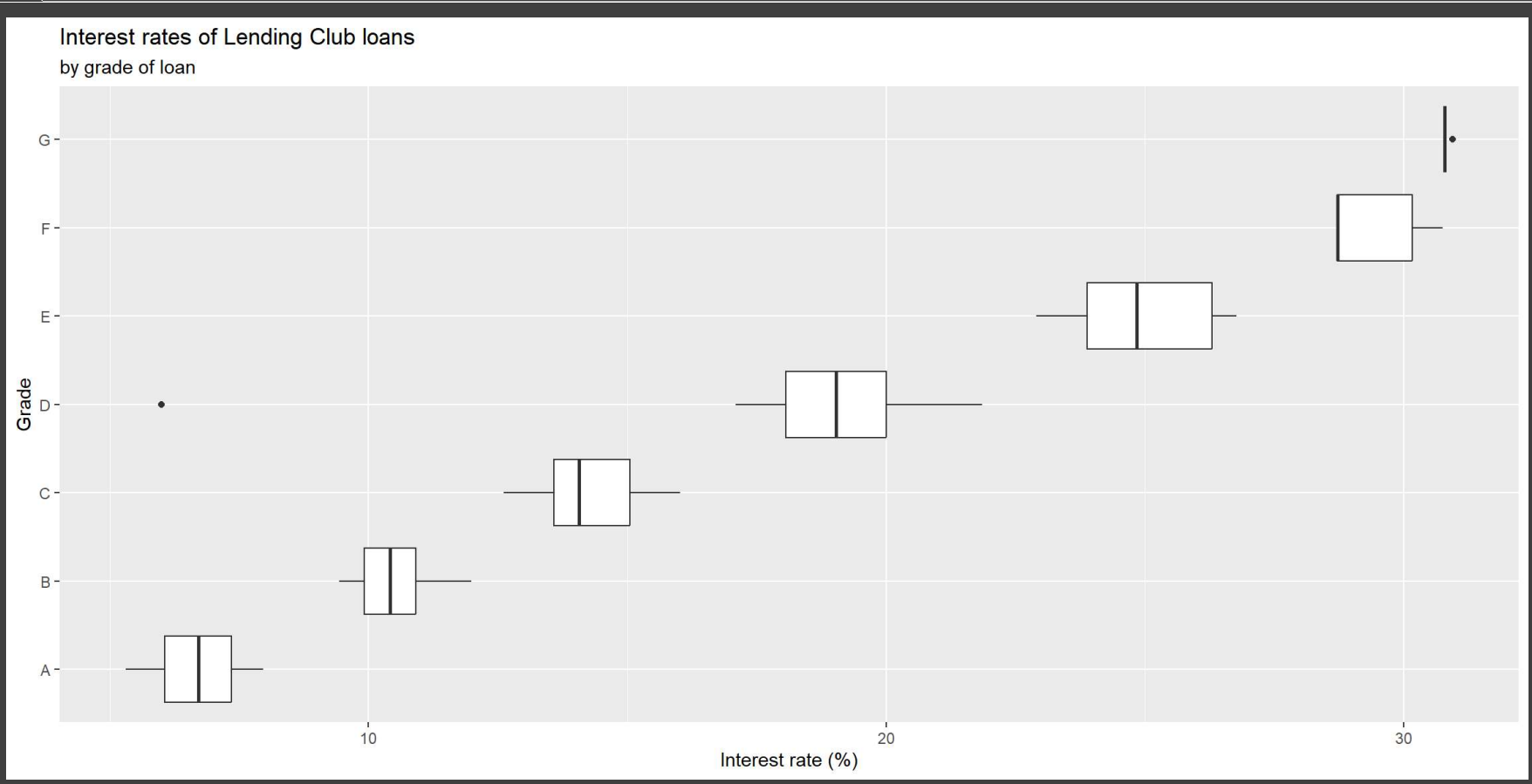
Box plot et outliers

```
1 ggplot(loans, aes(x = annual_income)) +
2   geom_boxplot()
```



Combinaison avec des variables qualitatives

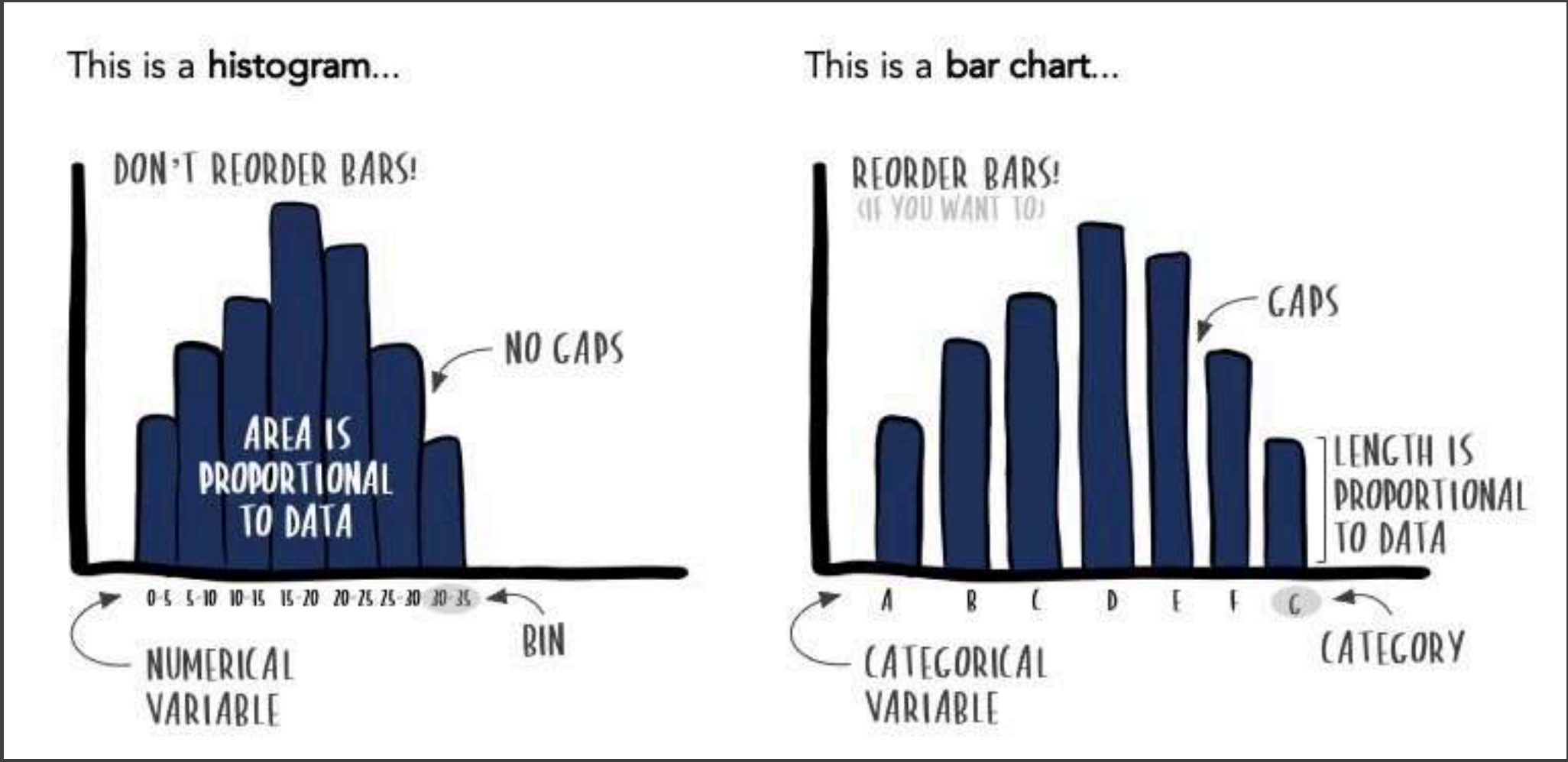
Plot | Code



Données qualitatives

Graphes en bar

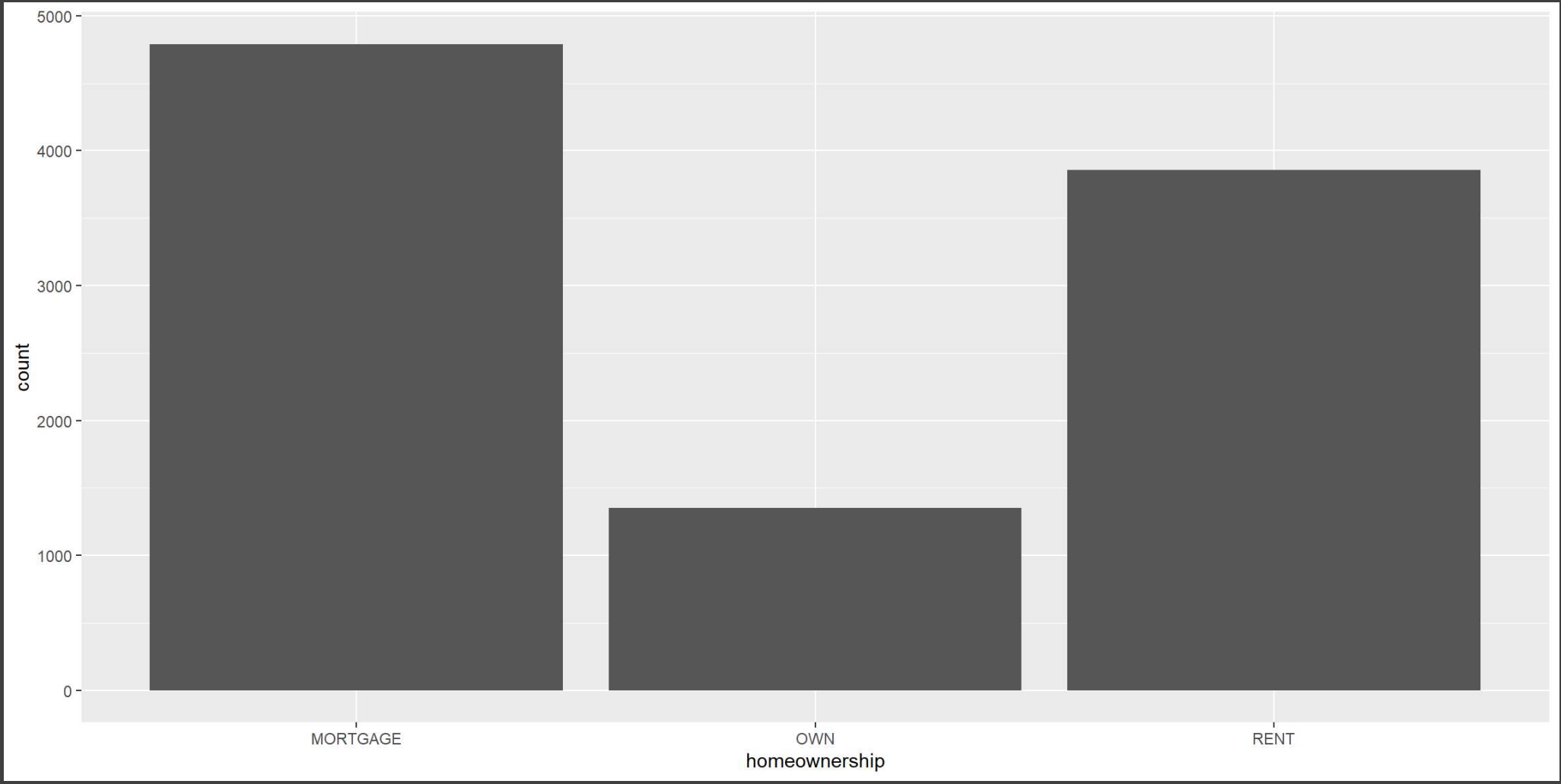
Histogramme vs bar charts



Source

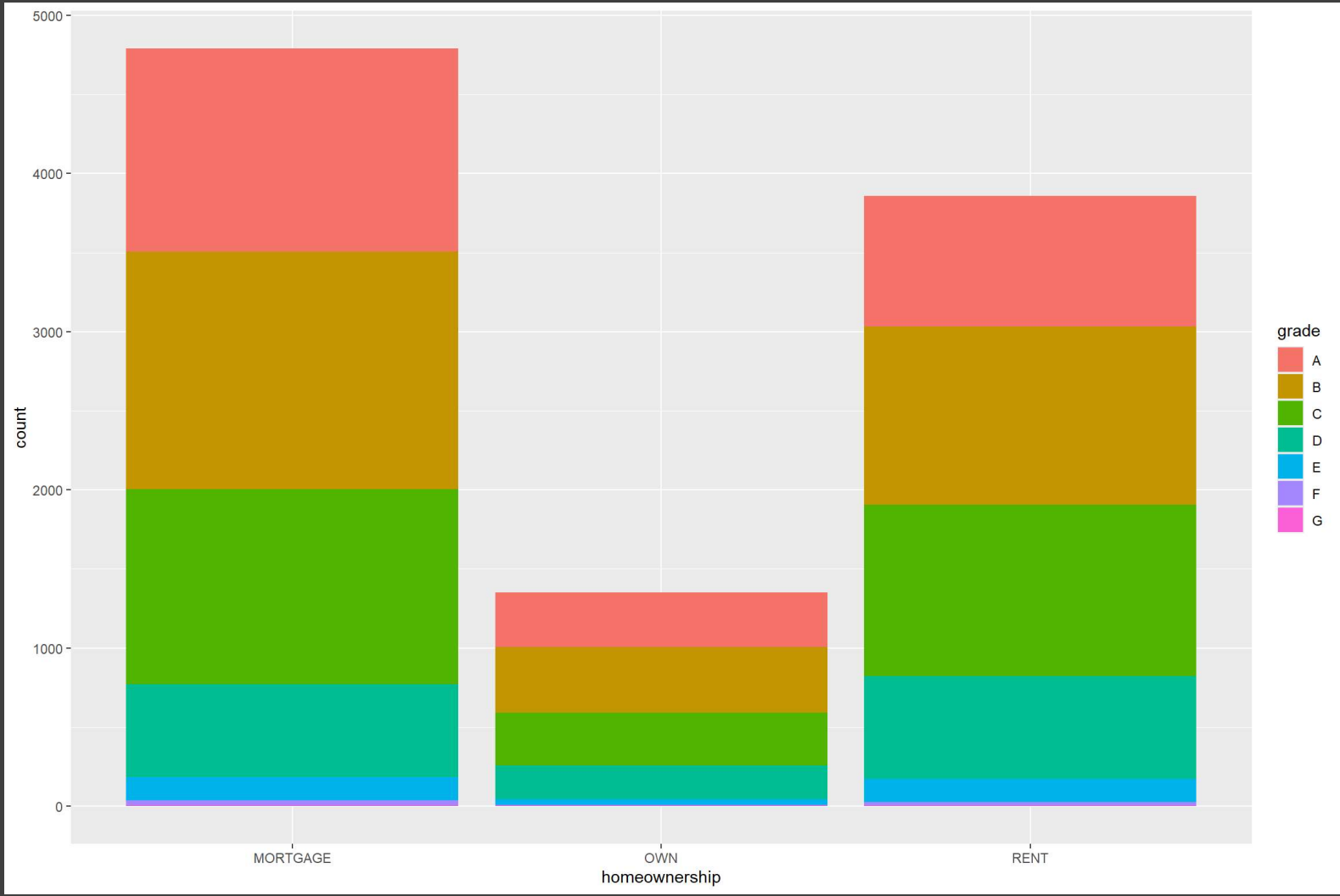
Bar chart

```
1 ggplot(loans, aes(x = homeownership)) +
2   geom_bar()
```



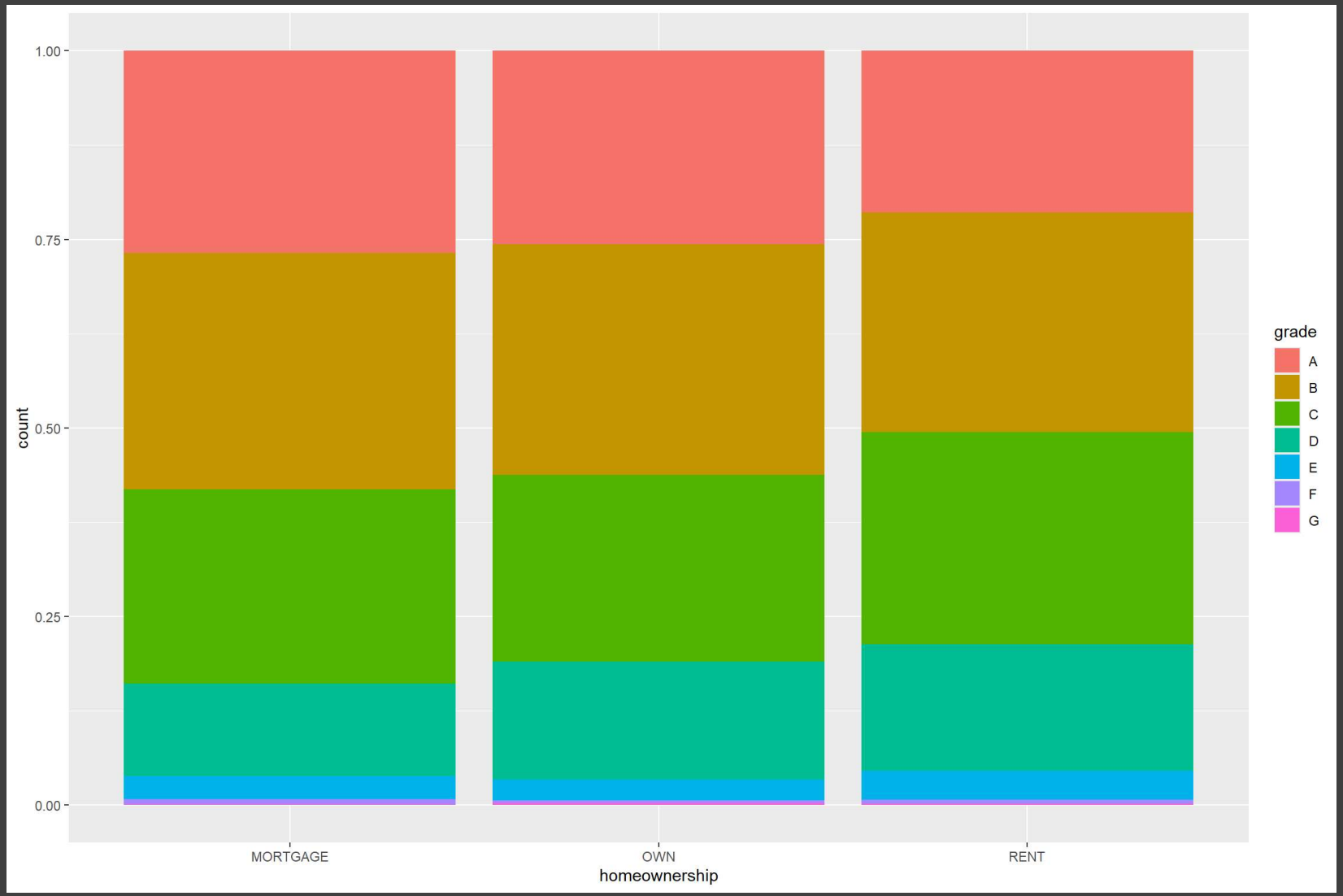
Bar chart segmenté

```
1 ggplot(loans, aes(x = homeownership,
2                   fill = grade)) +
3   geom_bar()
```

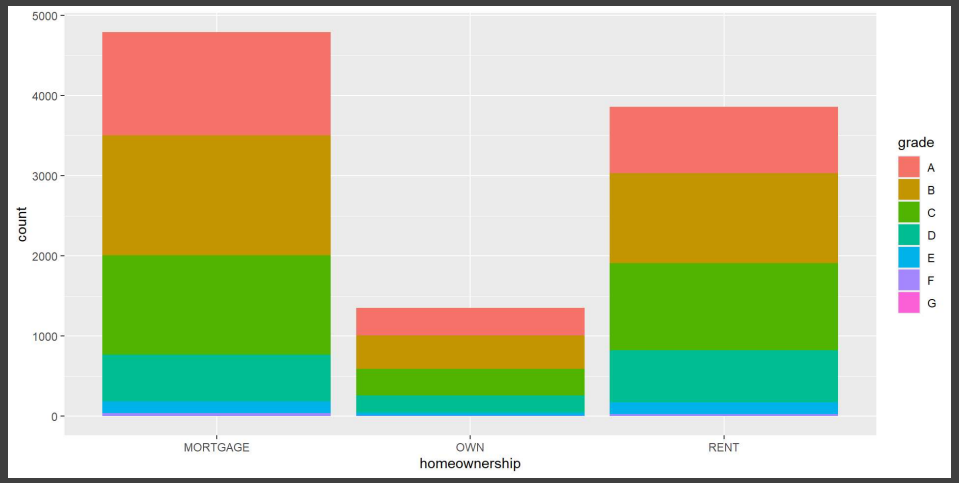


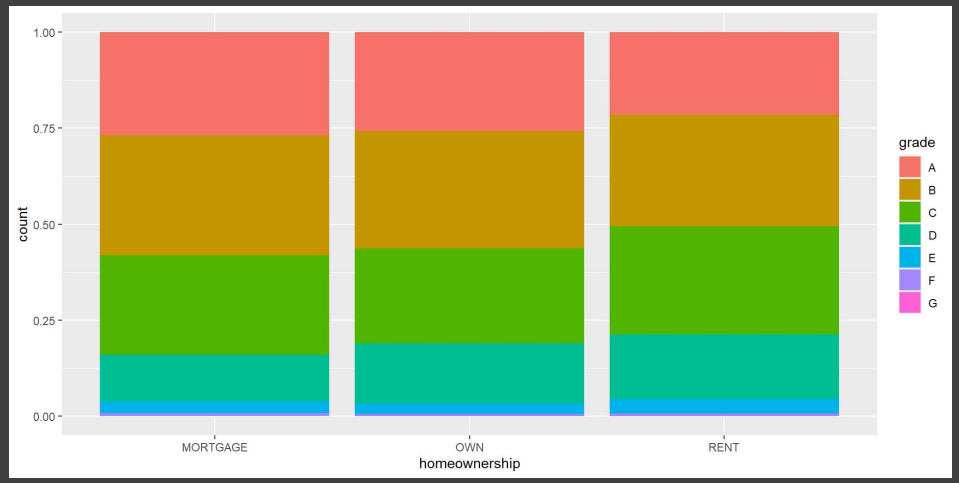
Bar chart segmenté

```
1 ggplot(loans, aes(x = homeownership, fill = grade)) +
2   geom_bar(position = "fill")
```



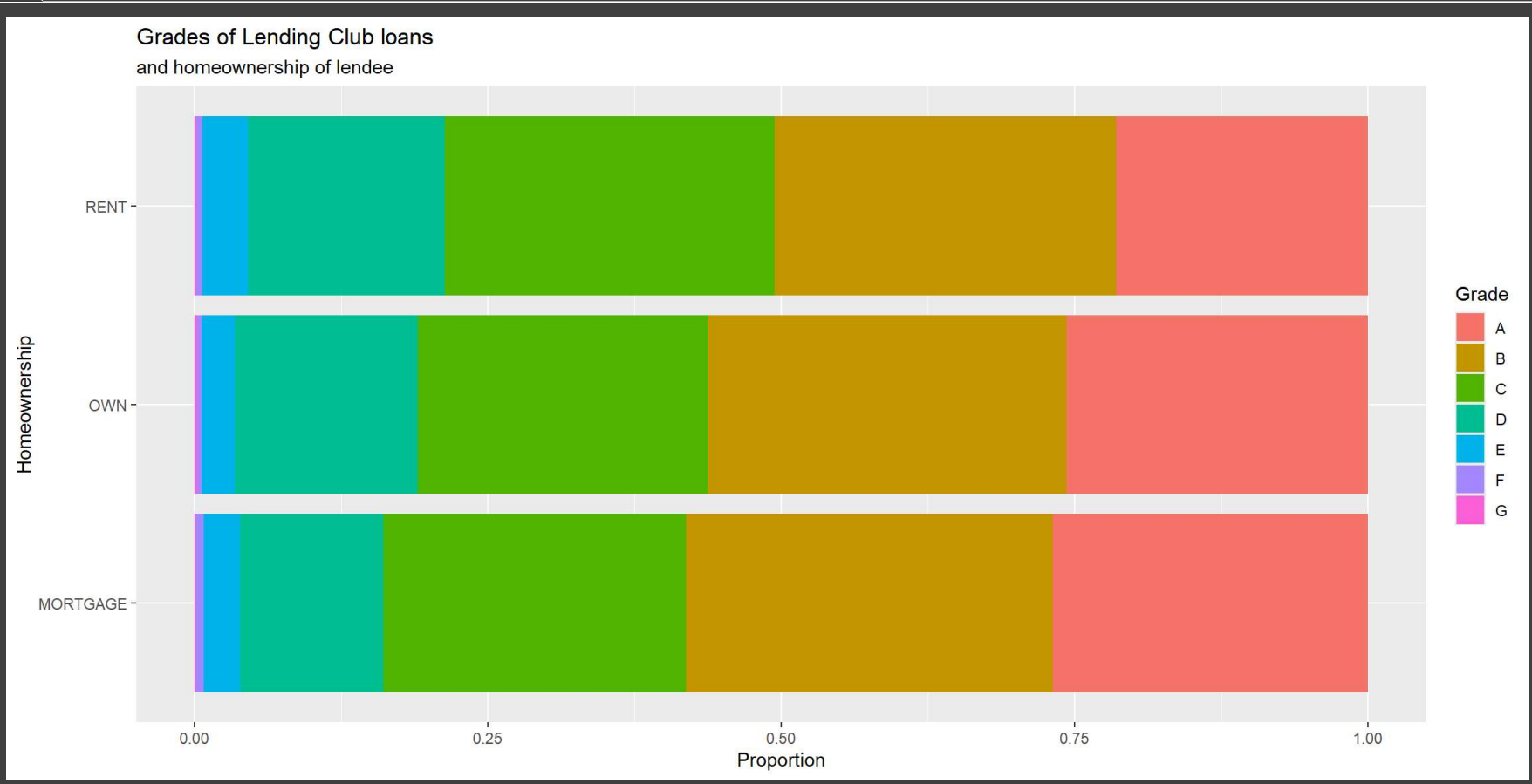
Lequel des deux graphes est plus adapté pour montrer la relation entre le fait d'être propriétaire et les notes de prêt ?





Bar plot horizontal

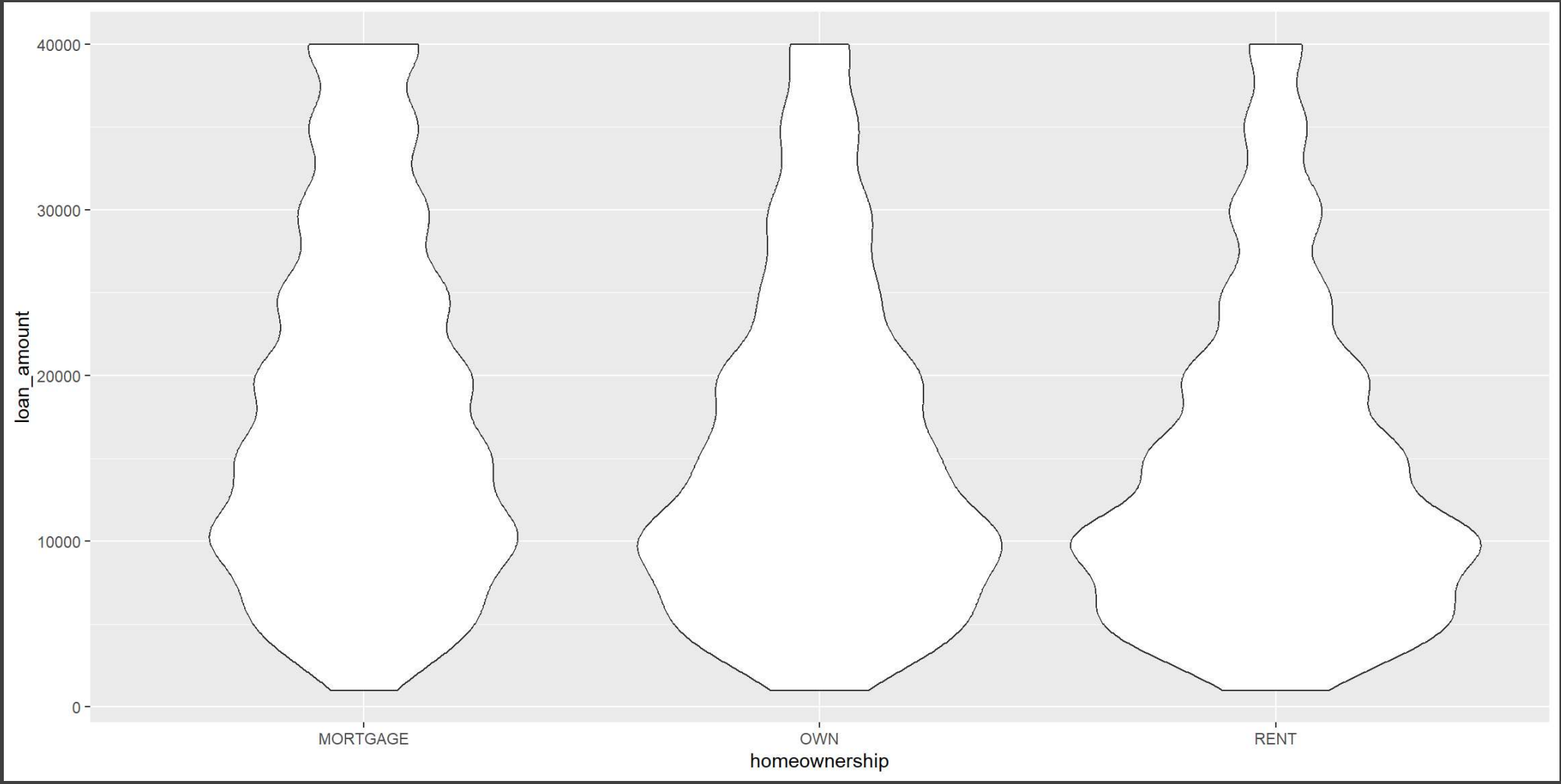
Plot | Code



Qualitative vs Quantitative

Violin plots

```
1 ggplot(loans, aes(x = homeownership, y = loan_amount)) +
2   geom_violin()
```



Ridge plots

```
1 library(ggridges)
2 ggplot(loans, aes(x = loan_amount, y = grade, fill = grade, color = grade)) +
3   geom_density_ridges(alpha = 0.5)
```