German Credit Data Exploration_2

Dr. Prashant Mishra

3/26/2018

1. Import the Clean data

We already have a clean file “german_credit_full.csv” to import.

```
credit_dataset <- read.csv("german_credit_full.csv",stringsAsFactors = TRUE)
str(credit_dataset)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ Class                    : Factor w/ 2 levels "Bad","Good": 2 1 2 2 1 2 2 2 2 1 ...
##  $ CheckingAccountStatus    : Factor w/ 4 levels "0.to.200","gt.200",..: 3 1 4 3 3 4 4 1 4 1 ...
##  $ Duration                 : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ CreditHistory            : Factor w/ 5 levels "Critical","Delay",..: 1 4 1 4 2 4 4 4 4 1 ...
##  $ Purpose                  : Factor w/ 10 levels "Business","DomesticAppliance",..: 7 7 3 4 5 3 4 10
##  $ Amount                   : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ SavingsAccountBonds      : Factor w/ 5 levels "100.to.500","500.to.1000",..: 5 4 4 4 4 5 2 4 3 4
##  $ EmploymentDuration       : Factor w/ 5 levels "0.to.1","1.to.4",..: 4 2 3 3 2 2 4 2 3 5 ...
##  $ InstallmentRatePercentage: int  4 2 2 2 3 2 3 2 2 4 ...
##  $ Personal                 : Factor w/ 4 levels "Female.NotSingle",..: 4 1 4 4 4 4 4 4 2 3 ...
##  $ OtherDebtorsGuarantors   : Factor w/ 3 levels "CoApplicant",..: 3 3 3 2 3 3 3 3 3 3 ...
##  $ ResidenceDuration        : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ Property                 : Factor w/ 4 levels "CarOther","Insurance",..: 3 3 3 2 4 4 2 1 3 1 ...
##  $ Age                      : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ OtherInstallmentPlans    : Factor w/ 3 levels "Bank","None",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Housing                  : Factor w/ 3 levels "ForFree","Own",..: 2 2 2 1 1 1 2 3 2 2 ...
##  $ NumberExistingCredits    : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ Job                      : Factor w/ 4 levels "Management.SelfEmp.HighlyQualified",..: 2 2 4 2 2 4
##  $ NumberPeopleMaintenance  : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ Telephone                : int  1 0 0 0 0 1 0 1 0 0 ...
##  $ ForeignWorker            : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
colnames(credit_dataset)
```

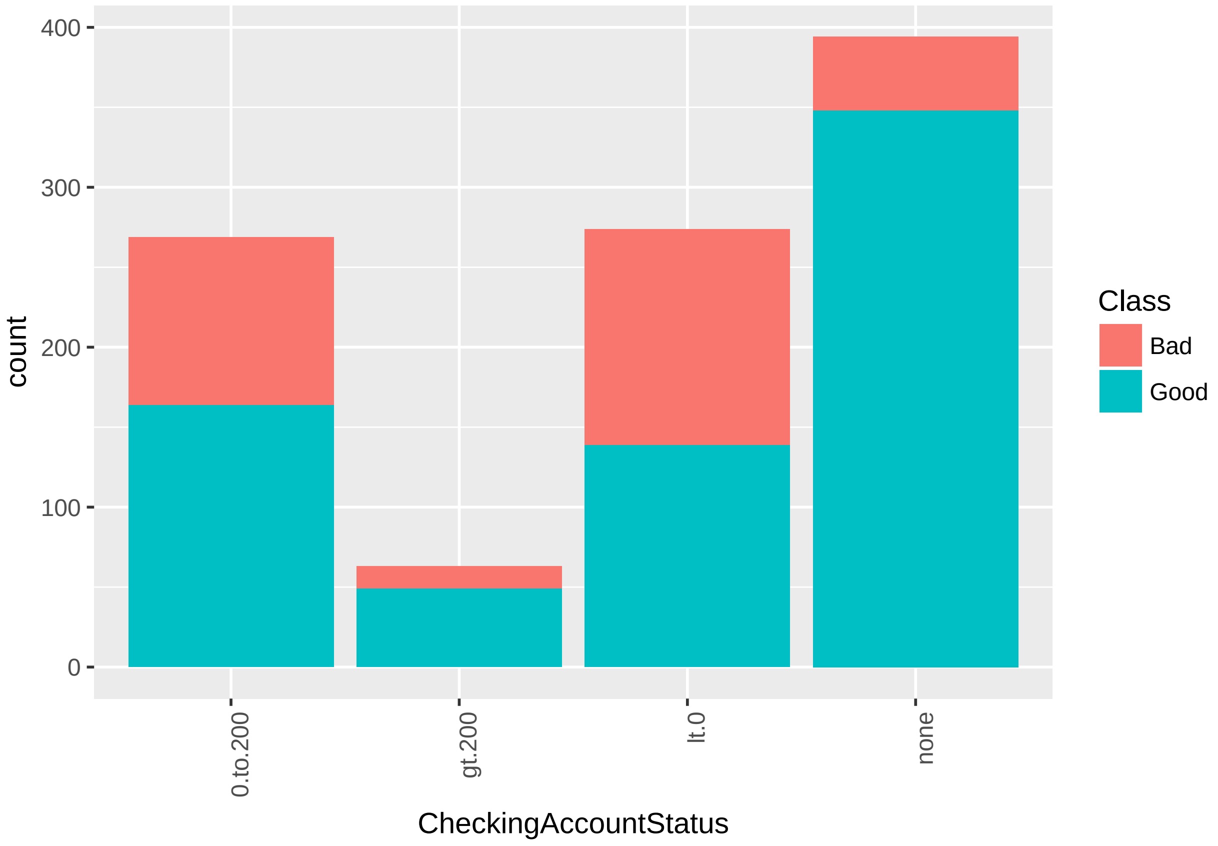
```
##  [1] "Class"                     "CheckingAccountStatus"
##  [3] "Duration"                  "CreditHistory"
##  [5] "Purpose"                   "Amount"
##  [7] "SavingsAccountBonds"       "EmploymentDuration"
##  [9] "InstallmentRatePercentage" "Personal"
## [11] "OtherDebtorsGuarantors"    "ResidenceDuration"
## [13] "Property"                  "Age"
## [15] "OtherInstallmentPlans"     "Housing"
## [17] "NumberExistingCredits"     "Job"
## [19] "NumberPeopleMaintenance"   "Telephone"
## [21] "ForeignWorker"
```

2. Explore Class vs Checking Account Status, Credit History and Employment Duration

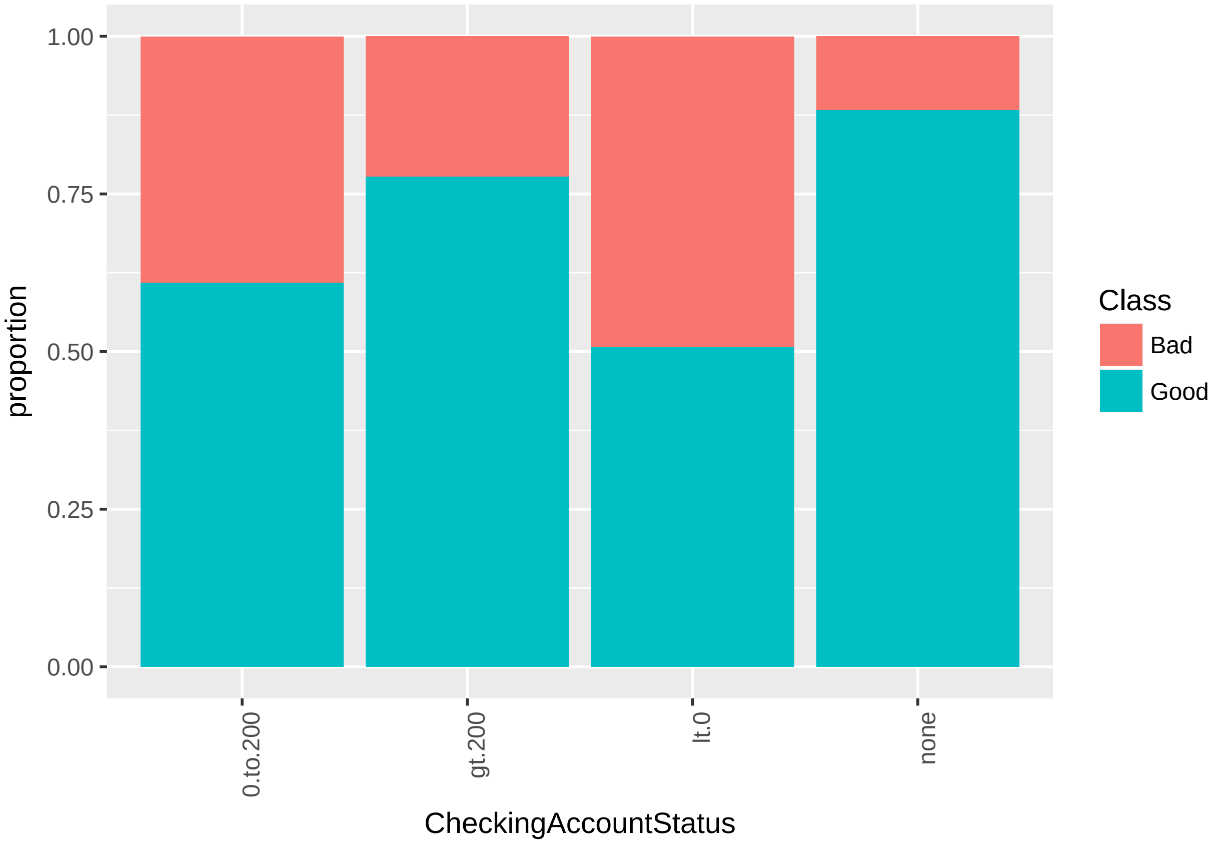
```r
library(gmodels)
CrossTable(credit_dataset$CheckingAccountStatus, credit_dataset$Class)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1000 
## 
## 
##                                      | credit_dataset$Class 
## credit_dataset$CheckingAccountStatus |       Bad |      Good | Row Total | 
## -------------------------------------|-----------|-----------|-----------|
##                             0.to.200 |       105 |       164 |       269 | 
##                                      |     7.317 |     3.136 |           | 
##                                      |     0.390 |     0.610 |     0.269 | 
##                                      |     0.350 |     0.234 |           | 
##                                      |     0.105 |     0.164 |           | 
## -------------------------------------|-----------|-----------|-----------|
##                               gt.200 |        14 |        49 |        63 | 
##                                      |     1.270 |     0.544 |           | 
##                                      |     0.222 |     0.778 |     0.063 | 
##                                      |     0.047 |     0.070 |           | 
##                                      |     0.014 |     0.049 |           | 
## -------------------------------------|-----------|-----------|-----------|
##                                 lt.0 |       135 |       139 |       274 | 
##                                      |    33.915 |    14.535 |           | 
##                                      |     0.493 |     0.507 |     0.274 | 
##                                      |     0.450 |     0.199 |           | 
##                                      |     0.135 |     0.139 |           | 
## -------------------------------------|-----------|-----------|-----------|
##                                 none |        46 |       348 |       394 | 
##                                      |    44.102 |    18.901 |           | 
##                                      |     0.117 |     0.883 |     0.394 | 
##                                      |     0.153 |     0.497 |           | 
##                                      |     0.046 |     0.348 |           | 
## -------------------------------------|-----------|-----------|-----------|
##                         Column Total |       300 |       700 |      1000 | 
##                                      |     0.300 |     0.700 |           | 
## -------------------------------------|-----------|-----------|-----------|
## 
## 
```

```
library(ggplot2)
pl1 = ggplot(credit_dataset, aes(x = CheckingAccountStatus, fill = Class));
pl2 = pl1 + geom_bar()
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



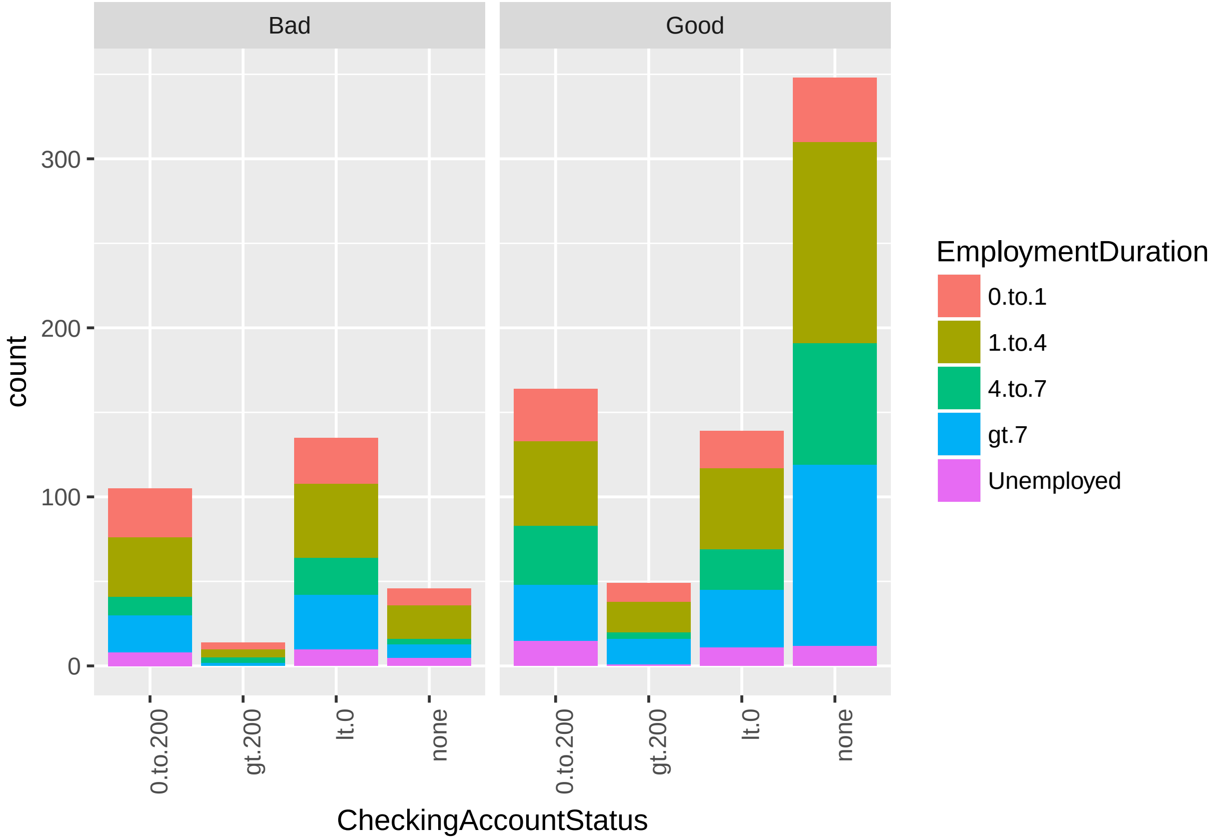
```
pl3 = pl1 + geom_bar(position = "fill") + ylab("proportion")
pl3+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



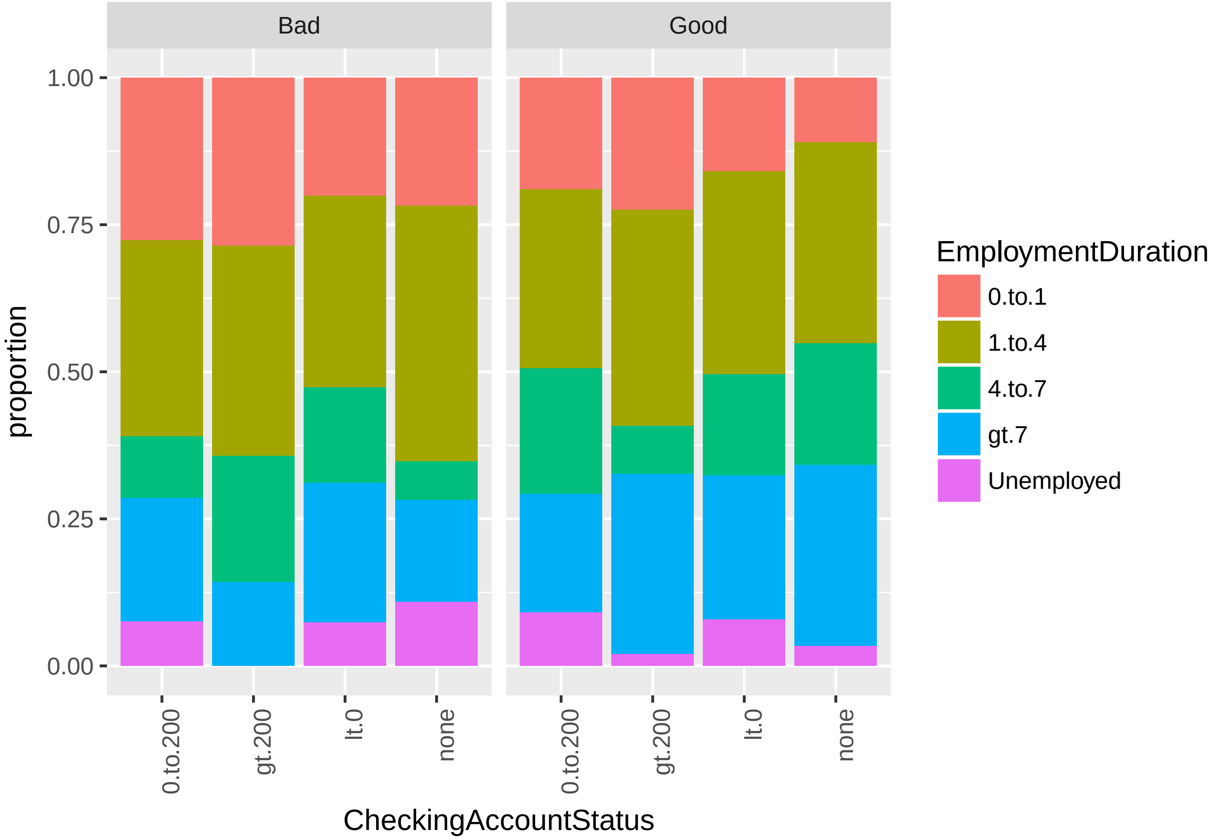
Two points:

1. It seems people who have Checking Account but only get amount between 0-200 DM are more likely to be a bad loan profile.
2. People who don't have a checking account might have account else where (other banks).

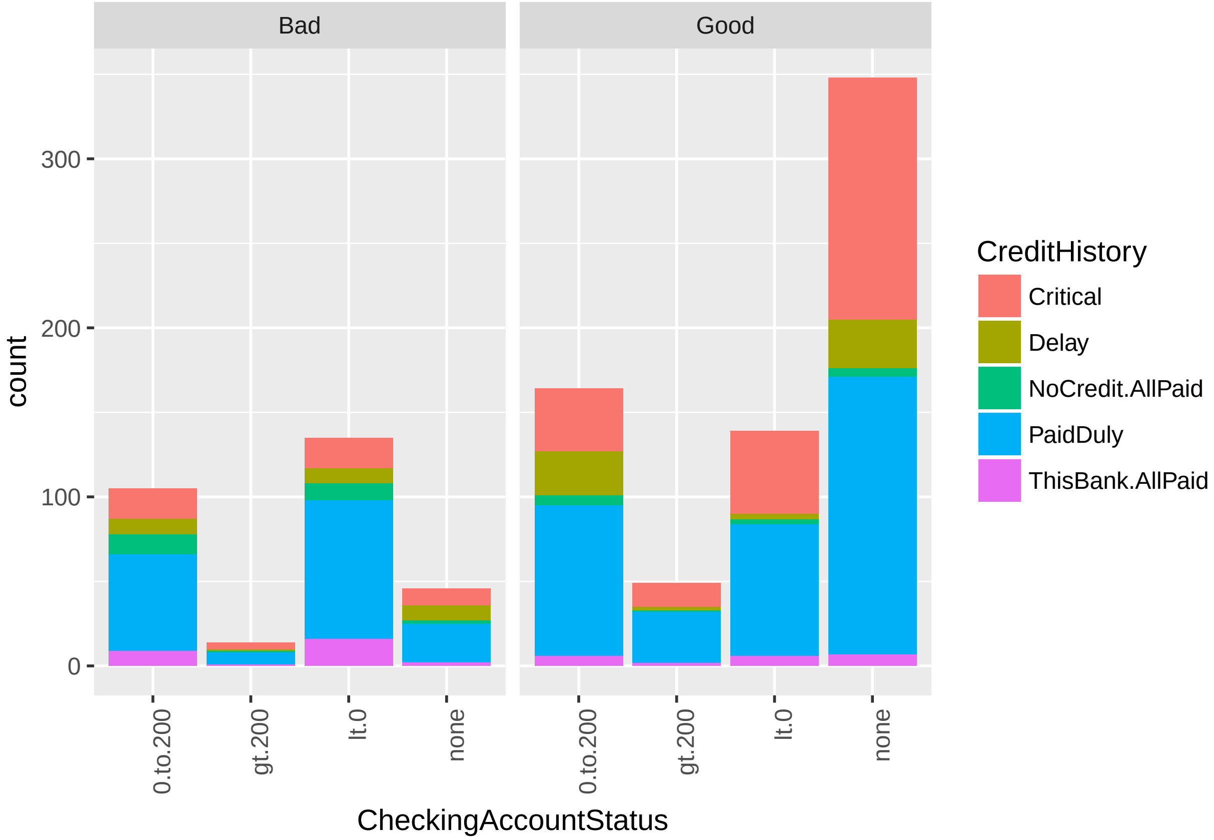
```
pl1 = ggplot(credit_dataset, aes(x = CheckingAccountStatus, fill = EmploymentDuration));
pl2 = pl1 + geom_bar()+facet_grid(~Class)
pl2+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



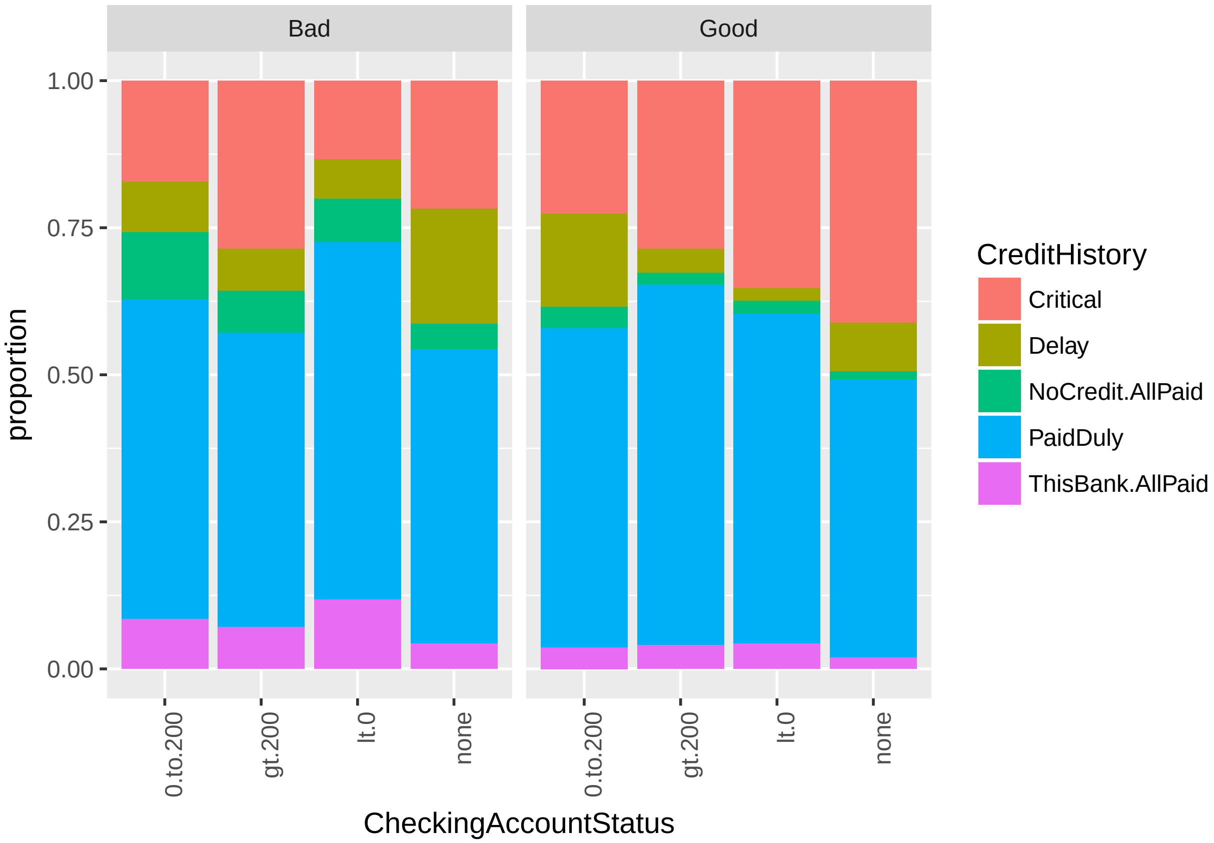
```
pl2 = pl1 + geom_bar(position = "fill") + ylab("proportion")+facet_grid(~Class)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
pl1 = ggplot(credit_dataset, aes(x = CheckingAccountStatus, fill = CreditHistory));
pl2 = pl1 + geom_bar()+facet_grid(~Class)
pl2+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
pl2 = pl1 + geom_bar(position = "fill") + ylab("proportion")+facet_grid(~Class)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



2. Explore Class vs Purpose, Personal and Property

```
CrossTable(credit_dataset$Purpose,credit_dataset$Class)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1000 
## 
## 
##                        | credit_dataset$Class 
## credit_dataset$Purpose |       Bad |      Good | Row Total | 
## -----------------------|-----------|-----------|-----------|
##               Business |        34 |        63 |        97 | 
##                        |     0.825 |     0.354 |           | 
##                        |     0.351 |     0.649 |     0.097 | 
##                        |     0.113 |     0.090 |           |
##                        |     0.034 |     0.063 |           |
## -----------------------|-----------|-----------|-----------|
##      DomesticAppliance |         4 |         8 |        12 |
##                        |     0.044 |     0.019 |           |
##                        |     0.333 |     0.667 |     0.012 |
##                        |     0.013 |     0.011 |           |
##                        |     0.004 |     0.008 |           |
## -----------------------|-----------|-----------|-----------|
##              Education |        22 |        28 |        50 |
##                        |     3.267 |     1.400 |           |
##                        |     0.440 |     0.560 |     0.050 |
##                        |     0.073 |     0.040 |           |
##                        |     0.022 |     0.028 |           |
## -----------------------|-----------|-----------|-----------|
##    Furniture.Equipment |        58 |       123 |       181 |
##                        |     0.252 |     0.108 |           |
##                        |     0.320 |     0.680 |     0.181 |
##                        |     0.193 |     0.176 |           |
##                        |     0.058 |     0.123 |           |
## -----------------------|-----------|-----------|-----------|
##                 NewCar |        89 |       145 |       234 |
##                        |     5.035 |     2.158 |           |
##                        |     0.380 |     0.620 |     0.234 |
##                        |     0.297 |     0.207 |           |
##                        |     0.089 |     0.145 |           |
## -----------------------|-----------|-----------|-----------|
##                 Others |         5 |         7 |        12 |
##                        |     0.544 |     0.233 |           |
##                        |     0.417 |     0.583 |     0.012 |
##                        |     0.017 |     0.010 |           |
##                        |     0.005 |     0.007 |           |
## -----------------------|-----------|-----------|-----------|
##       Radio.Television |        62 |       218 |       280 |
##                        |     5.762 |     2.469 |           |
##                        |     0.221 |     0.779 |     0.280 |
##                        |     0.207 |     0.311 |           |
##                        |     0.062 |     0.218 |           |
## -----------------------|-----------|-----------|-----------|
##                Repairs |         8 |        14 |        22 |
##                        |     0.297 |     0.127 |           |
##                        |     0.364 |     0.636 |     0.022 |
##                        |     0.027 |     0.020 |           |
##                        |     0.008 |     0.014 |           |
## -----------------------|-----------|-----------|-----------|
##             Retraining |         1 |         8 |         9 |
##                        |     1.070 |     0.459 |           |
##                        |     0.111 |     0.889 |     0.009 |
##                        |     0.003 |     0.011 |           |
##                        |     0.001 |     0.008 |           |
## -----------------------|-----------|-----------|-----------|
##                UsedCar |        17 |        86 |       103 |
##                        |     6.253 |     2.680 |           |
##                        |     0.165 |     0.835 |     0.103 |
##                        |     0.057 |     0.123 |           | 
##                        |     0.017 |     0.086 |           | 
## -----------------------|-----------|-----------|-----------|
##           Column Total |       300 |       700 |      1000 | 
##                        |     0.300 |     0.700 |           | 
## -----------------------|-----------|-----------|-----------|
## 
## 
```

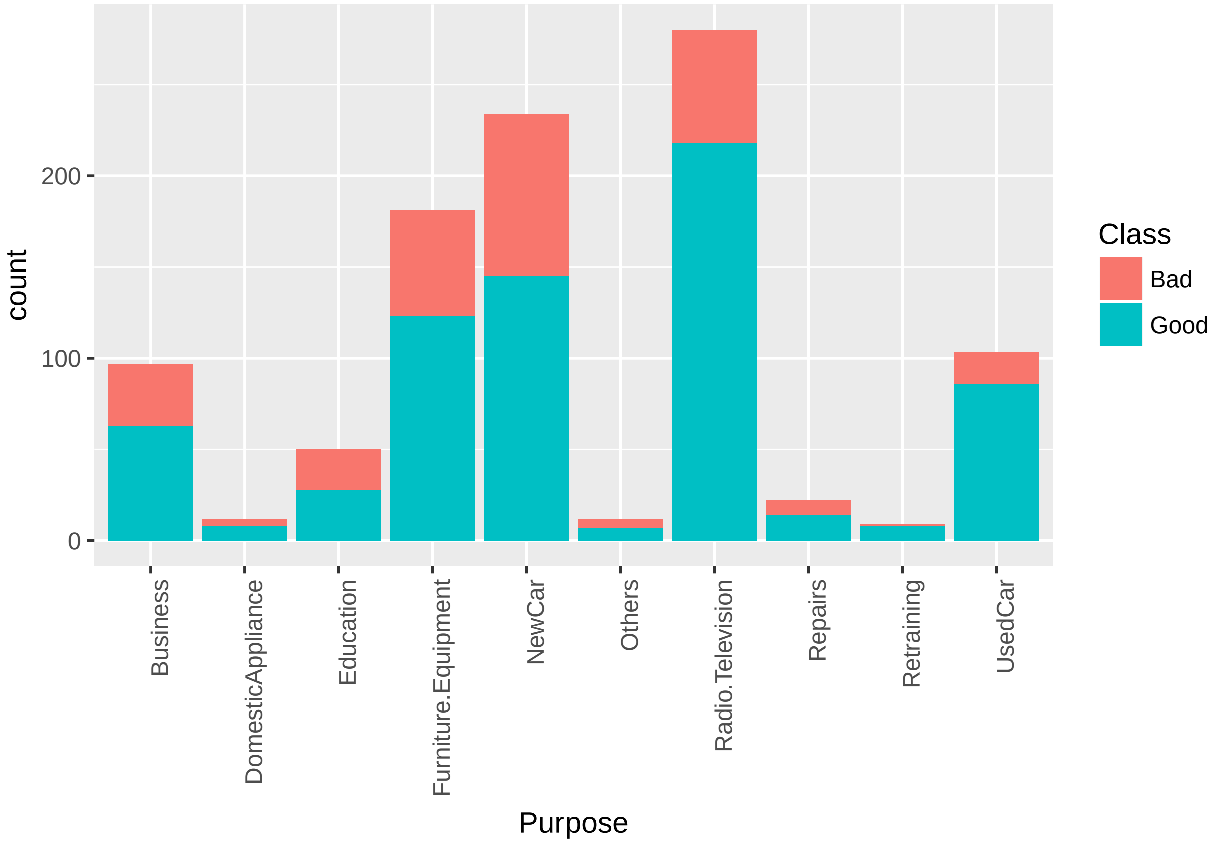
```
CrossTable(credit_dataset$Personal,credit_dataset$Class)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1000 
## 
## 
##                         | credit_dataset$Class 
## credit_dataset$Personal |       Bad |      Good | Row Total | 
## ------------------------|-----------|-----------|-----------|
##        Female.NotSingle |       109 |       201 |       310 | 
##                         |     2.753 |     1.180 |           | 
##                         |     0.352 |     0.648 |     0.310 | 
##                         |     0.363 |     0.287 |           | 
##                         |     0.109 |     0.201 |           | 
## ------------------------|-----------|-----------|-----------|
## Male.Divorced.Seperated |        20 |        30 |        50 | 
##                         |     1.667 |     0.714 |           | 
##                         |     0.400 |     0.600 |     0.050 | 
##                         |     0.067 |     0.043 |           | 
##                         |     0.020 |     0.030 |           | 
## ------------------------|-----------|-----------|-----------|
##    Male.Married.Widowed |        25 |        67 |        92 | 
##                         |     0.245 |     0.105 |           | 
##                         |     0.272 |     0.728 |     0.092 | 
##                         |     0.083 |     0.096 |           | 
##                         |     0.025 |     0.067 |           | 
## ------------------------|-----------|-----------|-----------|
##             Male.Single |       146 |       402 |       548 | 
##                         |     2.059 |     0.883 |           | 
##                         |     0.266 |     0.734 |     0.548 | 
##                         |     0.487 |     0.574 |           | 
##                         |     0.146 |     0.402 |           | 
## ------------------------|-----------|-----------|-----------|
##            Column Total |       300 |       700 |      1000 | 
##                         |     0.300 |     0.700 |           | 
## ------------------------|-----------|-----------|-----------|
##
##
```

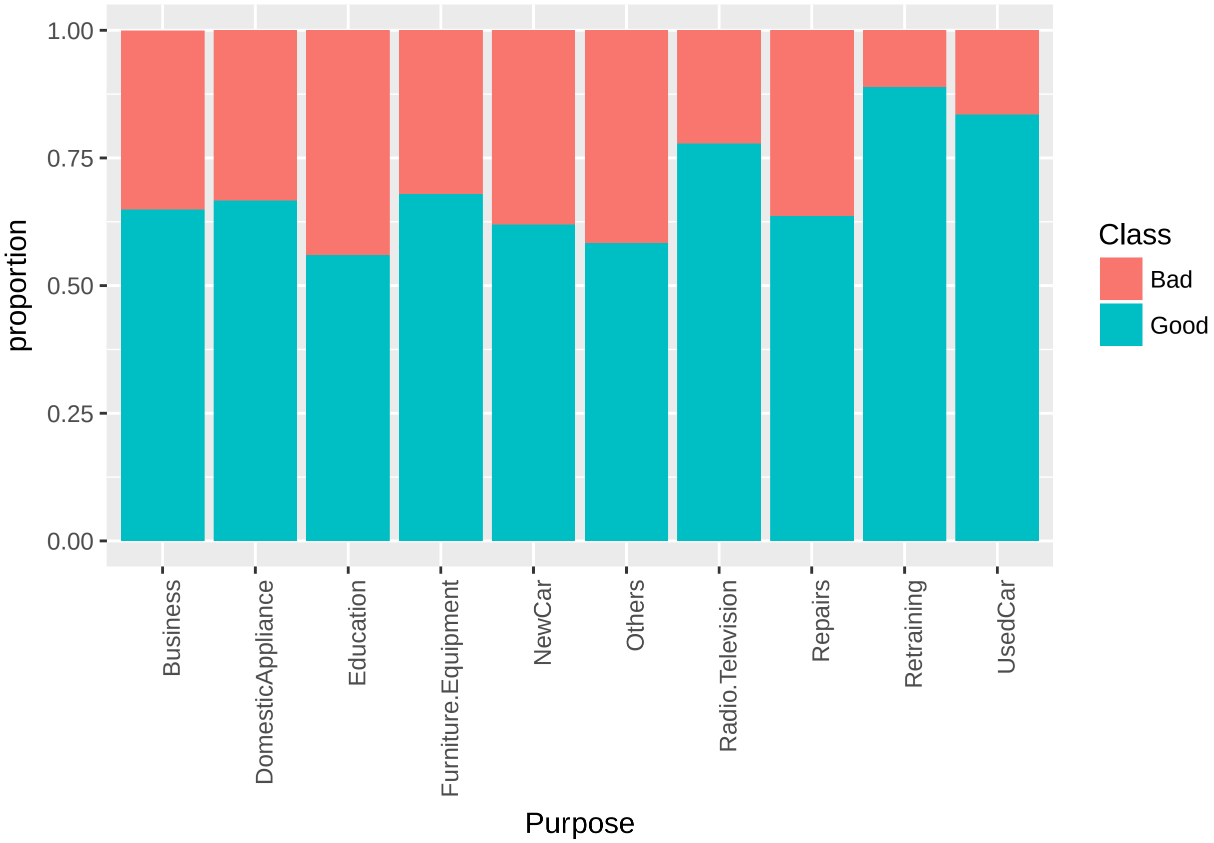
```
CrossTable(credit_dataset$Property,credit_dataset$Class)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1000
##
##
##                         | credit_dataset$Class
## credit_dataset$Property |       Bad |      Good | Row Total |
## ------------------------|-----------|-----------|-----------|
##                CarOther |       102 |       230 |       332 |
##                         |     0.058 |     0.025 |           |
##                         |     0.307 |     0.693 |     0.332 |
##                         |     0.340 |     0.329 |           |
##                         |     0.102 |     0.230 |           |
## ------------------------|-----------|-----------|-----------|
##               Insurance |        71 |       161 |       232 |
##                         |     0.028 |     0.012 |           |
##                         |     0.306 |     0.694 |     0.232 |
##                         |     0.237 |     0.230 |           |
##                         |     0.071 |     0.161 |           |
## ------------------------|-----------|-----------|-----------|
##              RealEstate |        60 |       222 |       282 |
##                         |     7.153 |     3.066 |           |
##                         |     0.213 |     0.787 |     0.282 |
##                         |     0.200 |     0.317 |           |
##                         |     0.060 |     0.222 |           |
## ------------------------|-----------|-----------|-----------|
##                 Unknown |        67 |        87 |       154 |
##                         |     9.365 |     4.013 |           |
##                         |     0.435 |     0.565 |     0.154 |
##                         |     0.223 |     0.124 |           |
##                         |     0.067 |     0.087 |           |
## ------------------------|-----------|-----------|-----------|
##            Column Total |       300 |       700 |      1000 |
##                         |     0.300 |     0.700 |           |
## ------------------------|-----------|-----------|-----------|
##
##
```

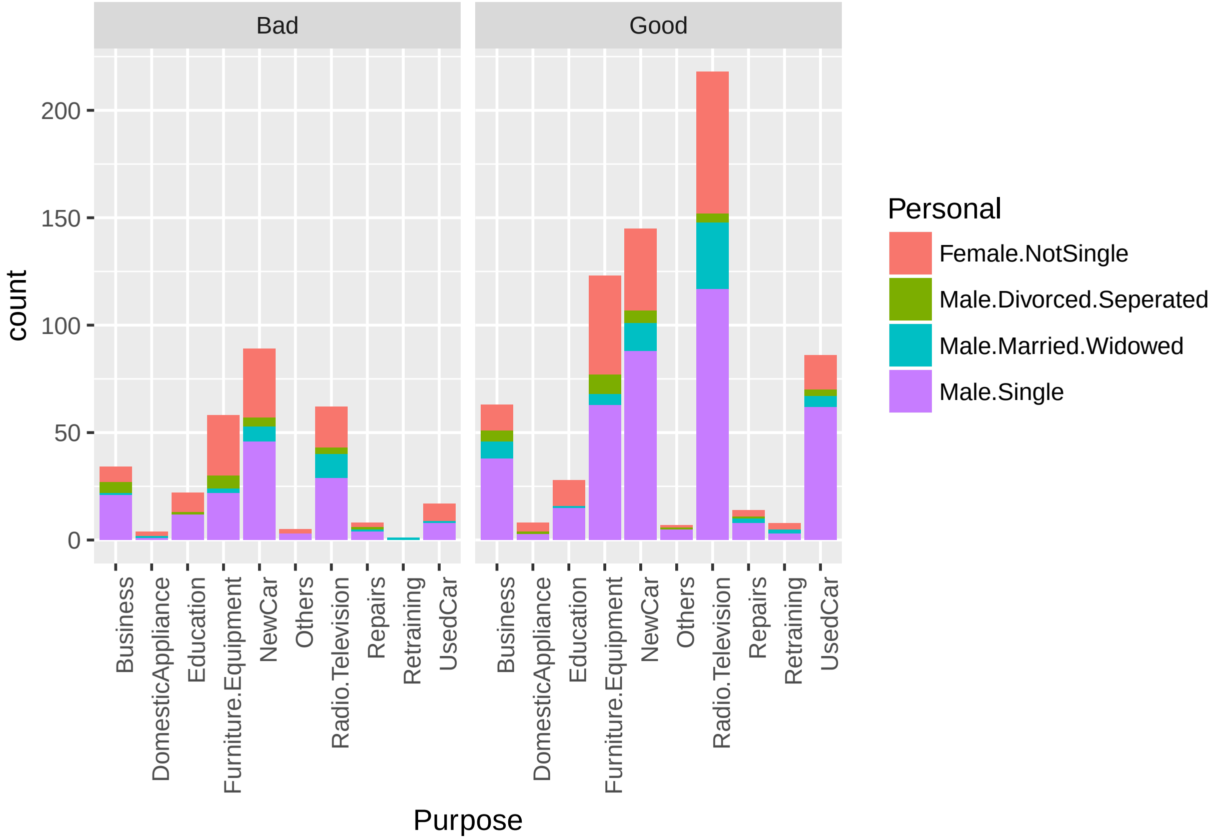
```
pl1 = ggplot(credit_dataset, aes(x = Purpose, fill = Class));
pl1 + geom_bar()+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



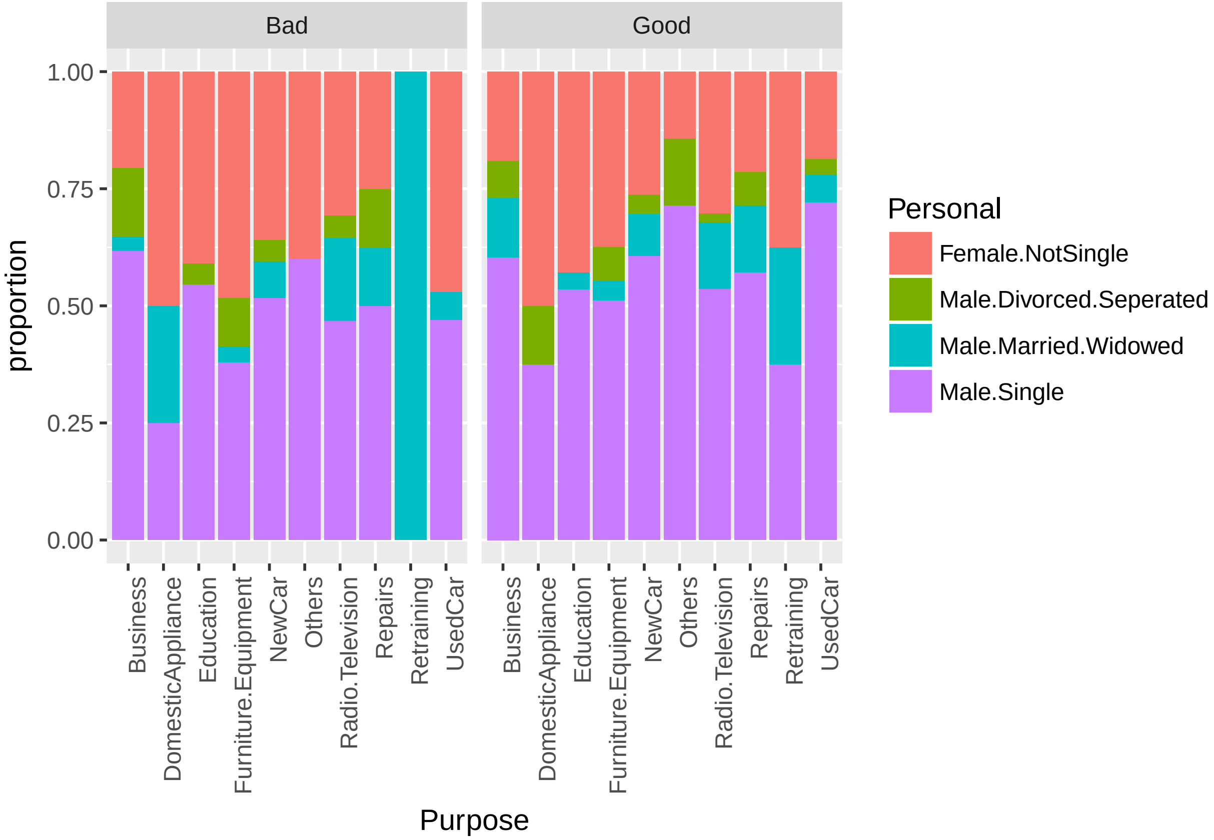
```
pl1 + geom_bar(position = "fill") + ylab("proportion")+ theme(axis.text.x = element_text(angle = 90, hju
```



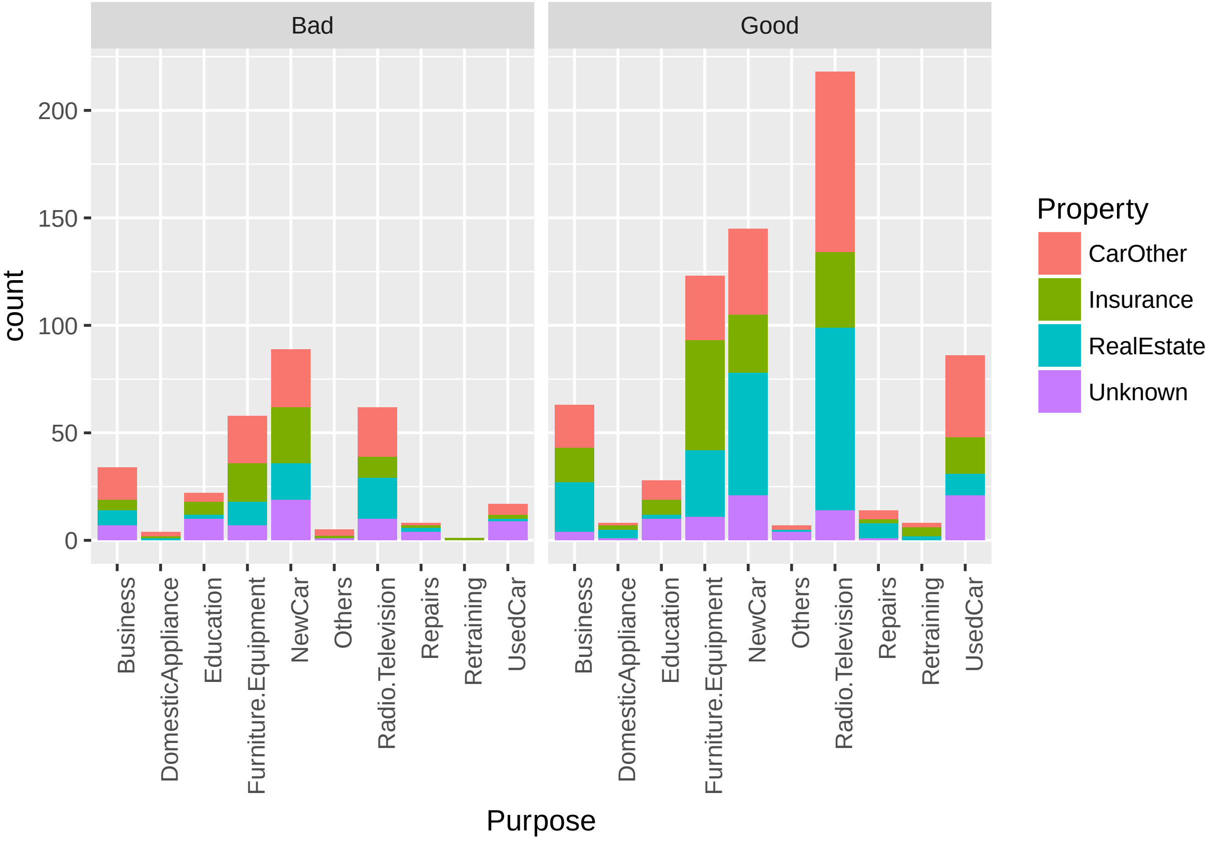
```
pl1 = ggplot(credit_dataset, aes(x = Purpose, fill = Personal));
pl2 = pl1 + geom_bar()+facet_grid(~Class)
pl2+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



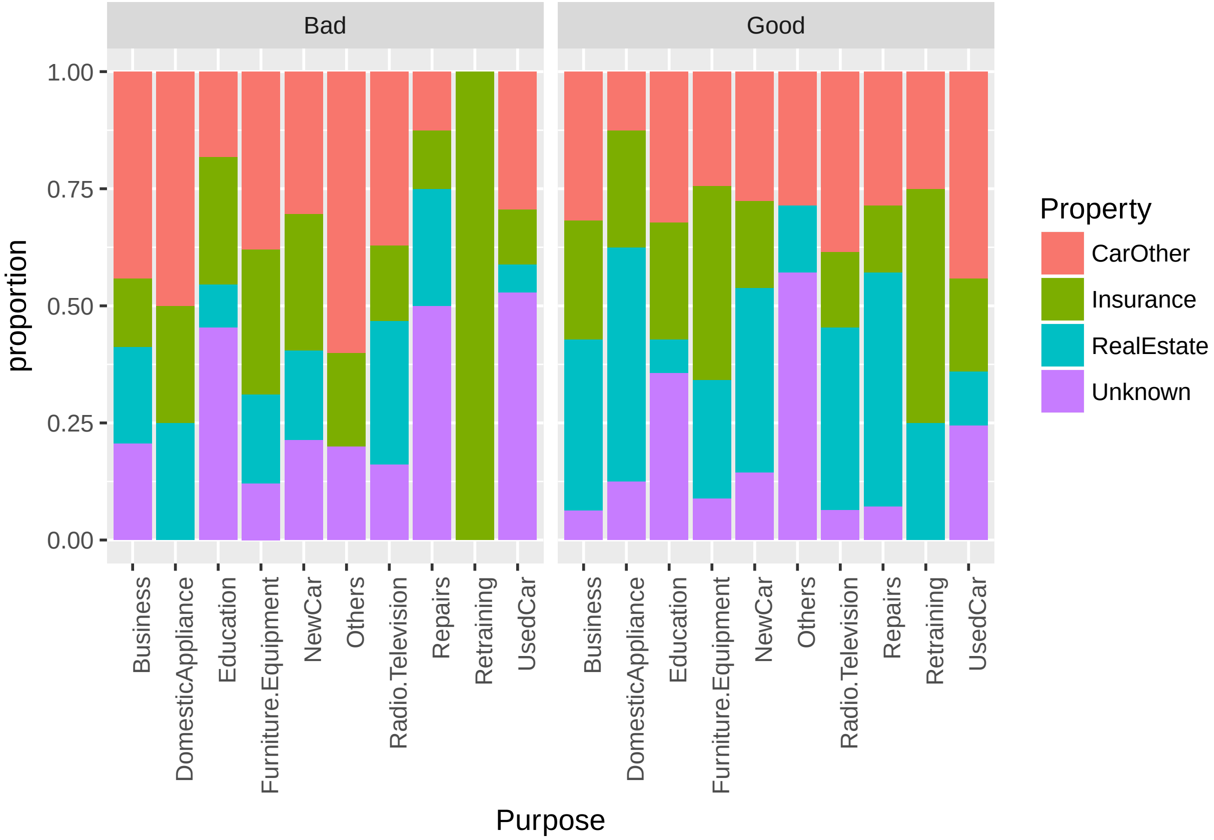
```
pl2 = pl1 + geom_bar(position = "fill") + ylab("proportion")+facet_grid(~Class)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
pl1 = ggplot(credit_dataset, aes(x = Purpose, fill = Property));
pl2 = pl1 + geom_bar()+facet_grid(~Class)
pl2+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

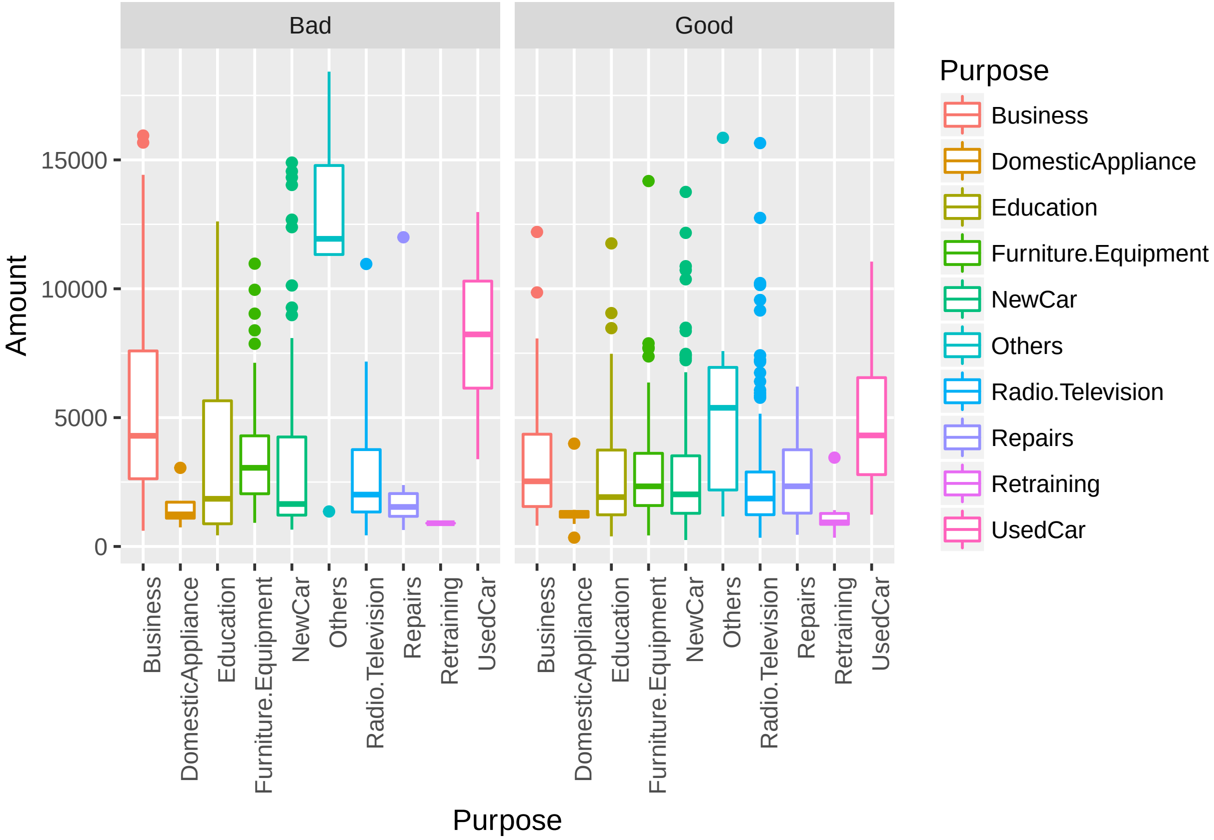


```
pl2 = pl1 + geom_bar(position = "fill") + ylab("proportion")+facet_grid(~Class)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

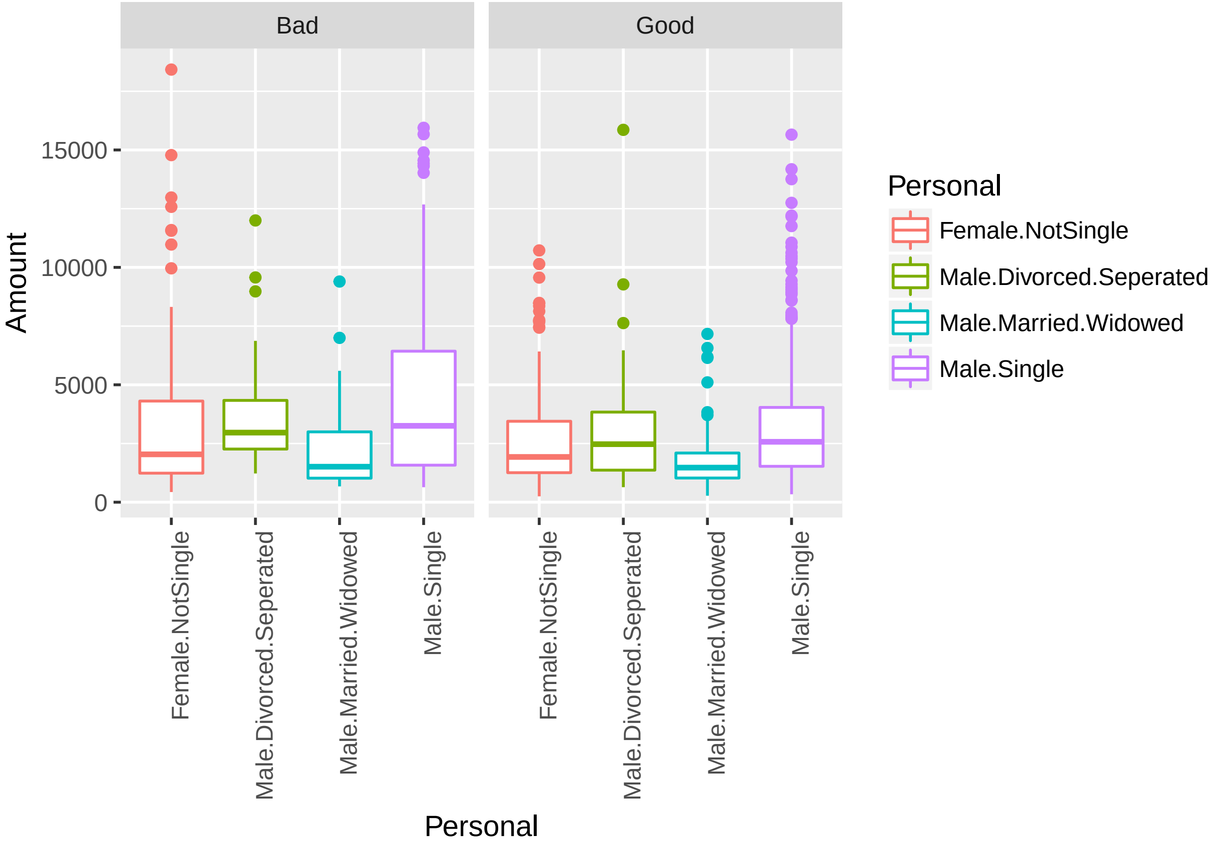


1. New Car loan seems to be worse than used car loans and any other loan.
2. Property doesn't seem to play a huge role as a deciding factor.

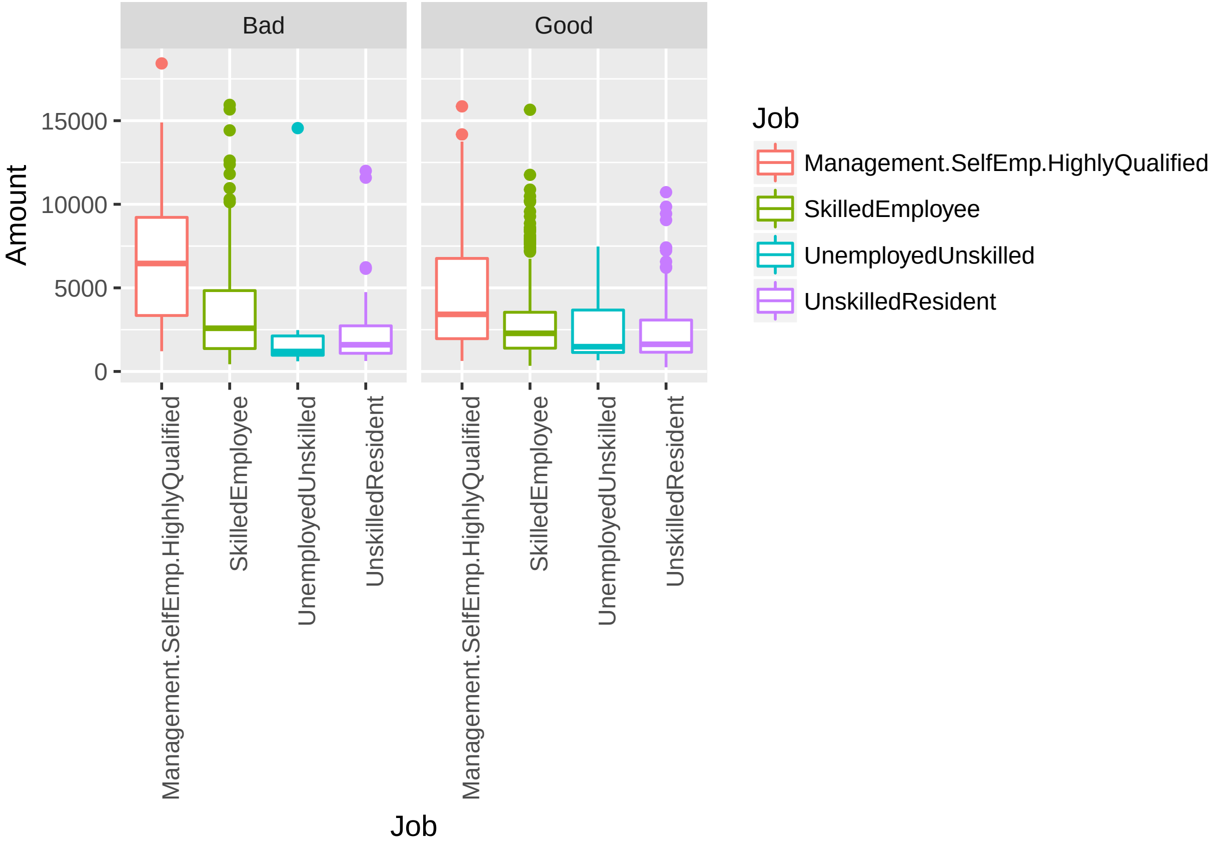
```
pl1 = ggplot(credit_dataset, aes(x = Purpose, y = Amount, color=Purpose));
pl2 = pl1 + geom_boxplot()+facet_grid(~Class)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



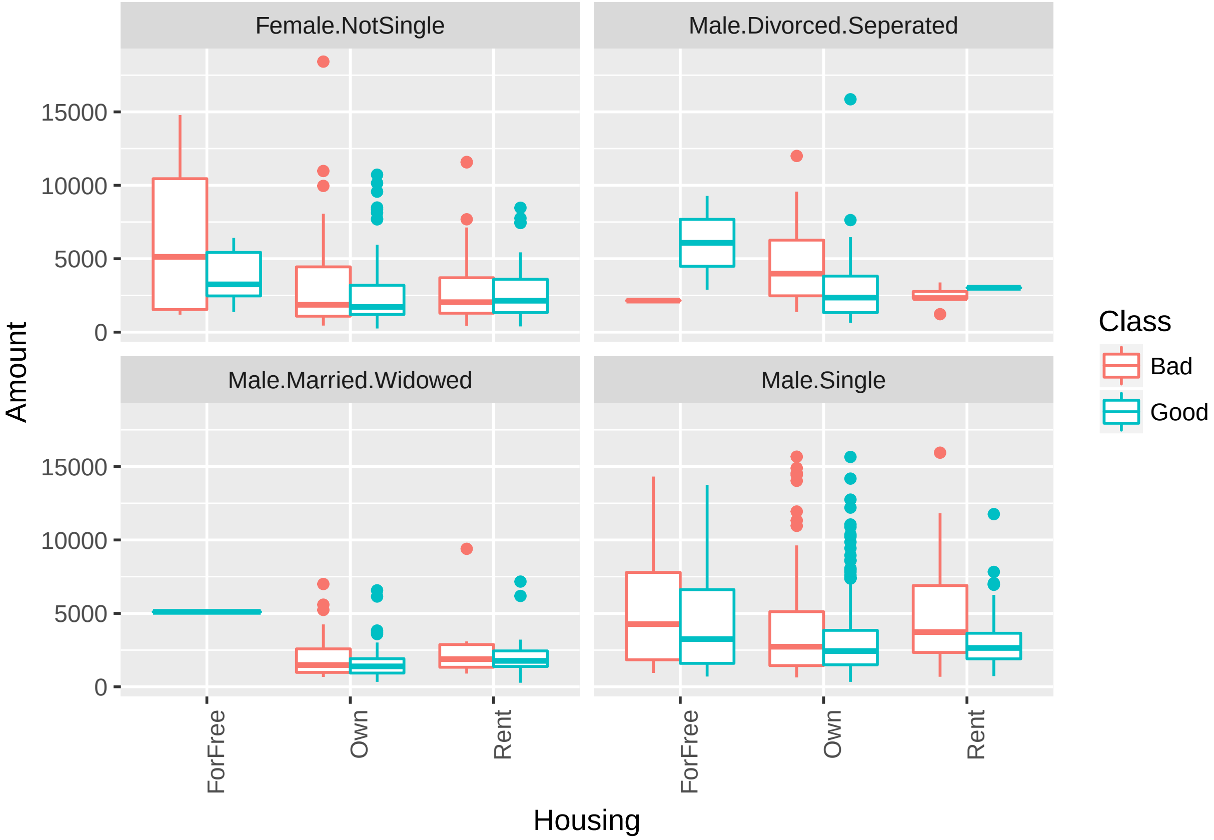
```
pl1 = ggplot(credit_dataset, aes(x = Personal, y = Amount, color=Personal));
pl2 = pl1 + geom_boxplot()+facet_grid(~Class)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



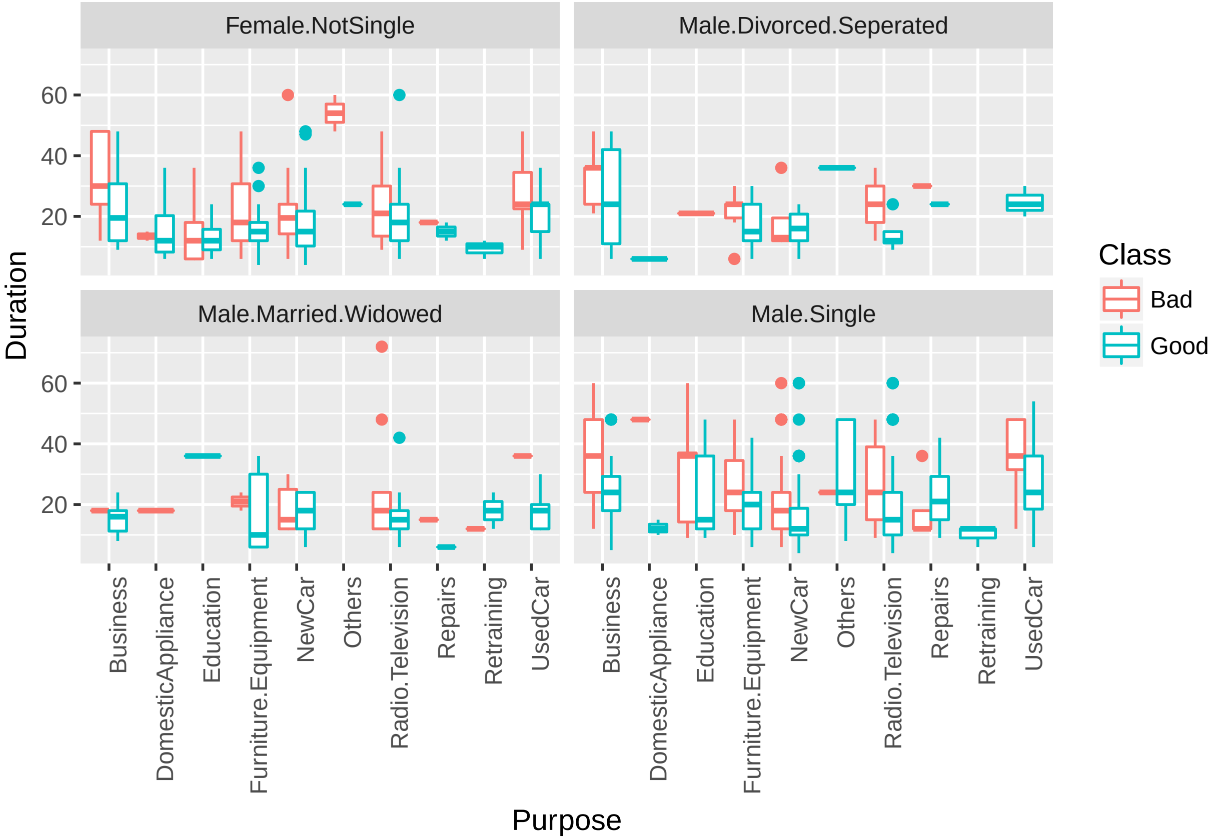
```
pl1 = ggplot(credit_dataset, aes(x = Job, y = Amount, color=Job));
pl2 = pl1 + geom_boxplot()+facet_wrap(~Class,ncol = 2)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



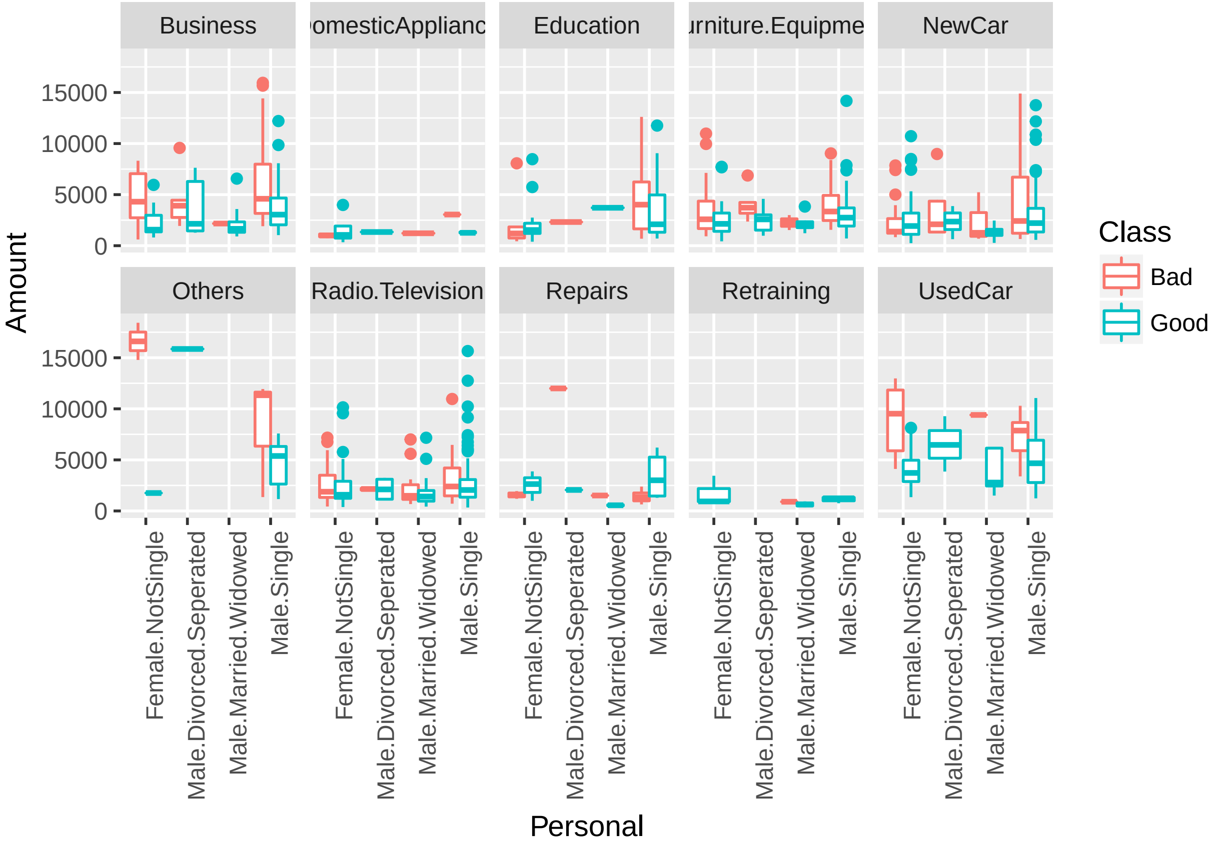
```
pl1 = ggplot(credit_dataset, aes(x = Housing, y = Amount, color=Class));
pl2 = pl1 + geom_boxplot()+facet_wrap(~Personal,ncol=2)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
pl1 = ggplot(credit_dataset, aes(x = Purpose, y = Duration, color=Class));
pl2 = pl1 + geom_boxplot()+facet_wrap(~Personal,ncol=2)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

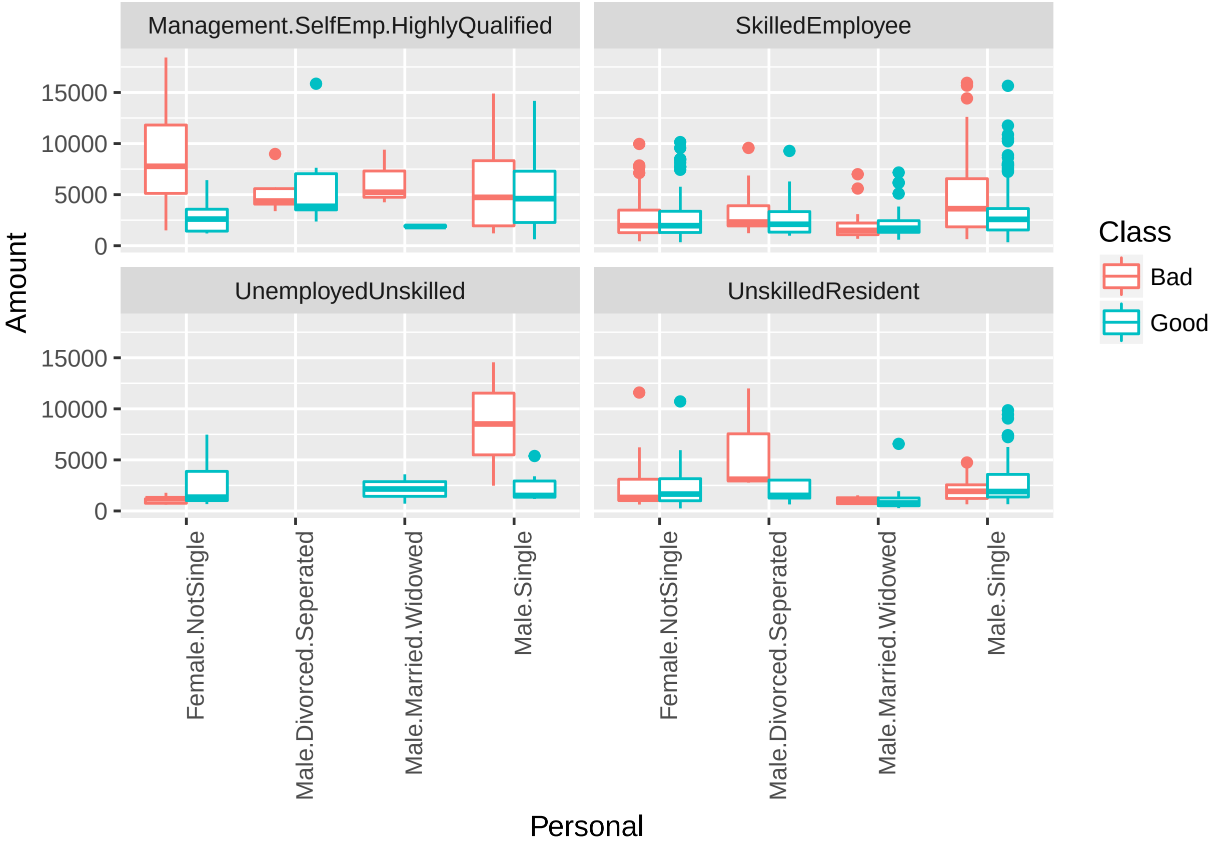


```
pl1 = ggplot(credit_dataset, aes(x = Personal, y = Amount, color=Class));
pl2 = pl1 + geom_boxplot()+facet_wrap(~Purpose, ncol = 5)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



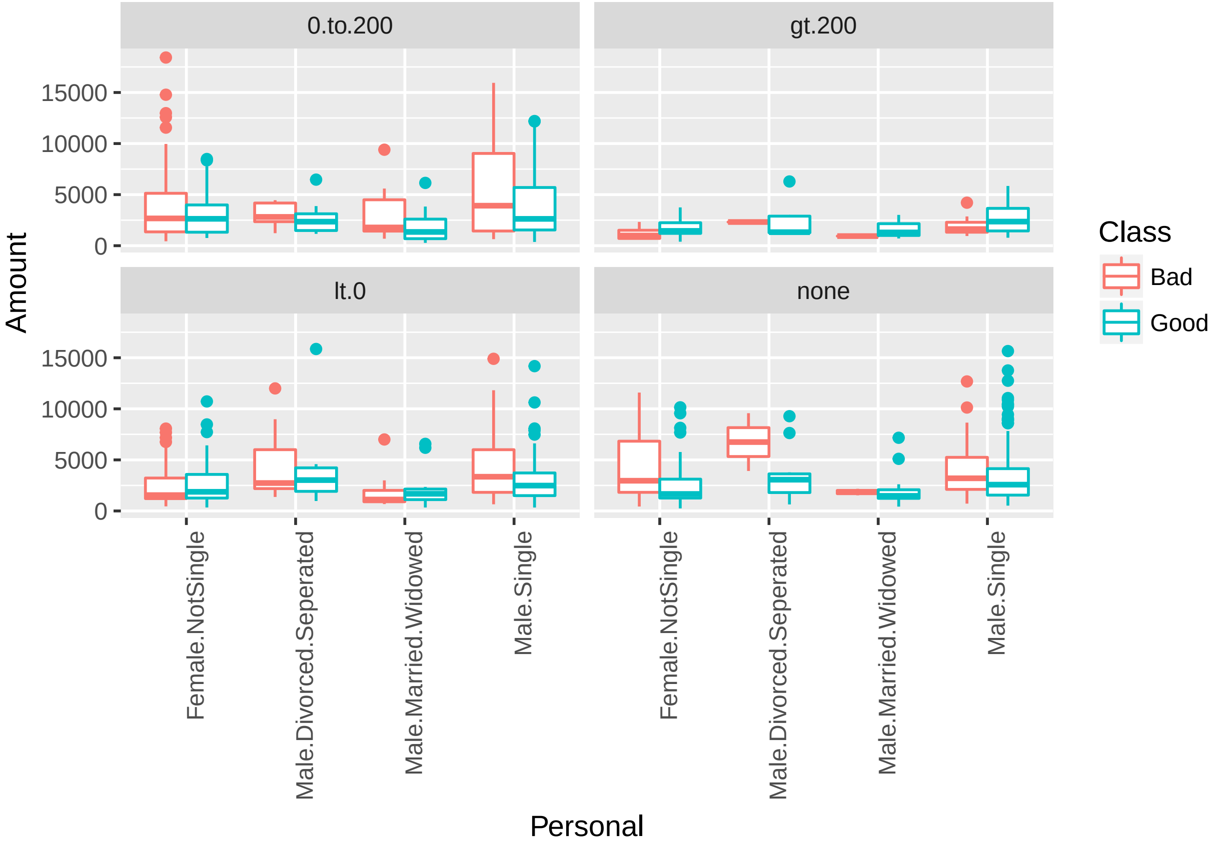
1. Used car data show that that Female.NotSingle tend to take higher loan for Used car, and are more likely bad loans.
2. Unsual data for Others, we might need to get more data on that.

```
pl1 = ggplot(credit_dataset, aes(x = Personal, y = Amount, color=Class));
pl2 = pl1 + geom_boxplot()+facet_wrap(~Job,ncol=2)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

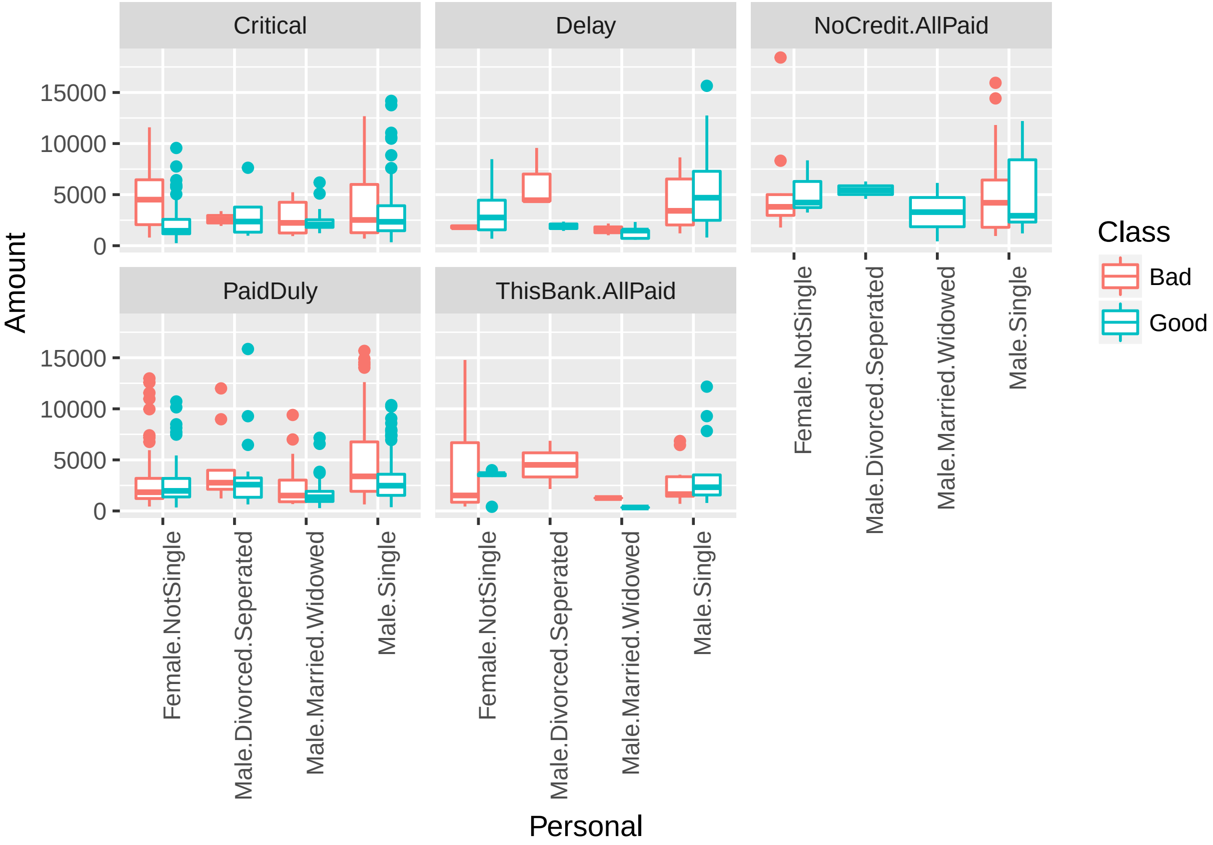


1. Male single, Unemployed and Unskilled— high risk 2. Female Not Single — Also high risk 3. Male Divored/Separated Unskilled Resident – high risk

```
pl1 = ggplot(credit_dataset, aes(x = Personal, y = Amount, color=Class));
pl2 = pl1 + geom_boxplot()+facet_wrap(~CheckingAccountStatus,ncol=2)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
pl1 = ggplot(credit_dataset, aes(x = Personal, y = Amount, color=Class));
pl2 = pl1 + geom_boxplot()+facet_wrap(~CreditHistory,ncol=3)
pl2 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Just some detective skill: We can see the outlier in NoCredit.AllPaid, Female.NotSingle. Some one from Highskilled, management or self employed category with checking account status between 0 to 200 took the loan for others category.