R Notebook

Code

Set up of the notebook :

Hide

```
#### Session Setup ----
rm(list = ls())
gc()
```

```
          used  (Mb) gc trigger  (Mb)  max used  (Mb)
Ncells  5632734 300.9   10451122 558.2  10451122 558.2
Vcells 18893692 144.2  114869471 876.4 125536623 957.8
```

Hide

```
set.seed(786)
Time = Sys.time()

list.of.packages <- c("tidyverse",
                      "readxl",
                      "writexl",
                      "lubridate",
                      "timetk",
                      "modeltime",
                      "tidymodels",
                      "purrr",
                      "h2o",
                      "DataExplorer",
                      "modeldata",
                      "tidyquant",
                      "plotly",
                      "caret")

new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"] )]
if(length(new.packages)) install.packages(new.packages, dependencies = TRUE)

for(i in list.of.packages){
  library(i, character.only = TRUE)
}
```

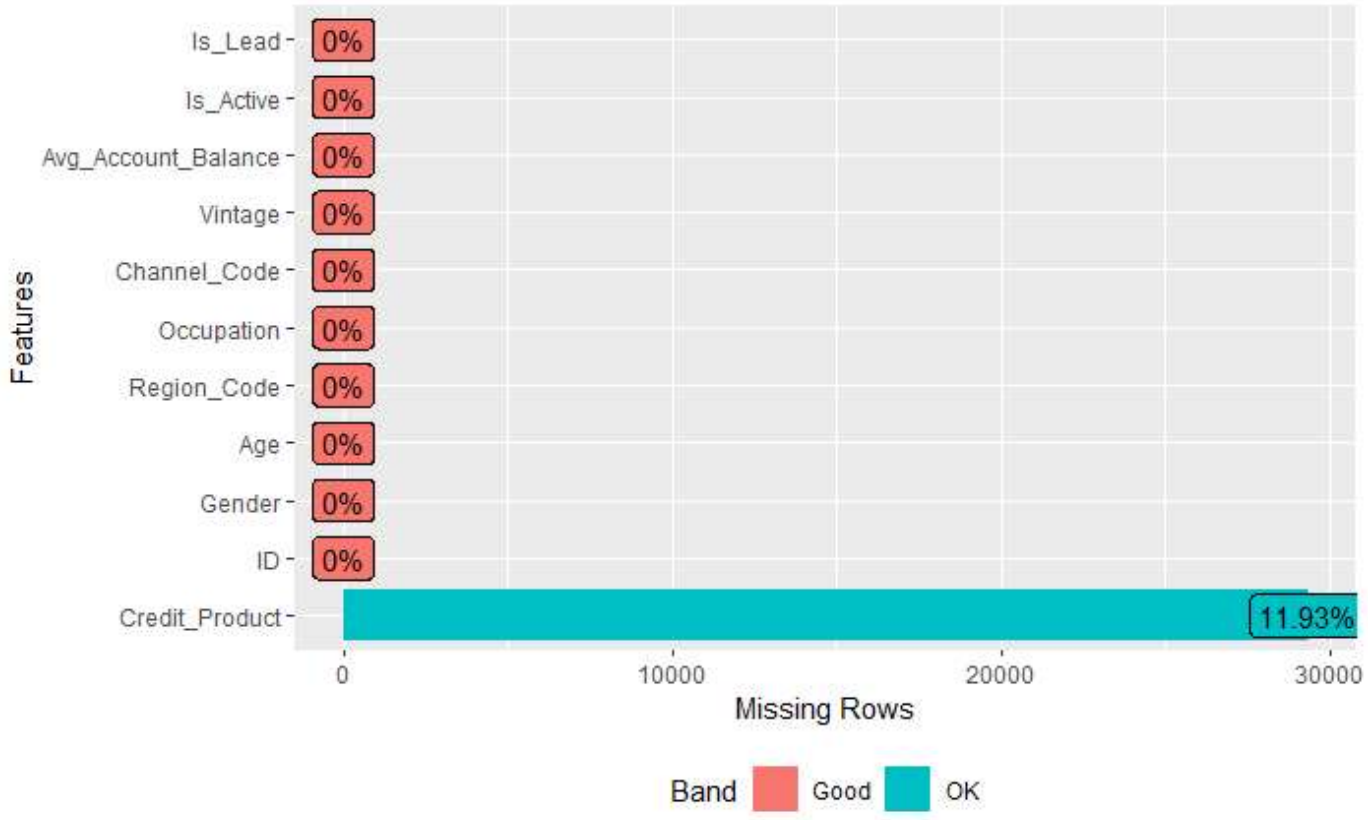
Hide

```
#### Read the data ----
train_tbl <- read_csv("../data/train_s3TEQDk.csv")
test_tbl <- read_csv("../data/test_mSzZ8RL.csv")
```

Exploratory Data Analysis :

Hide

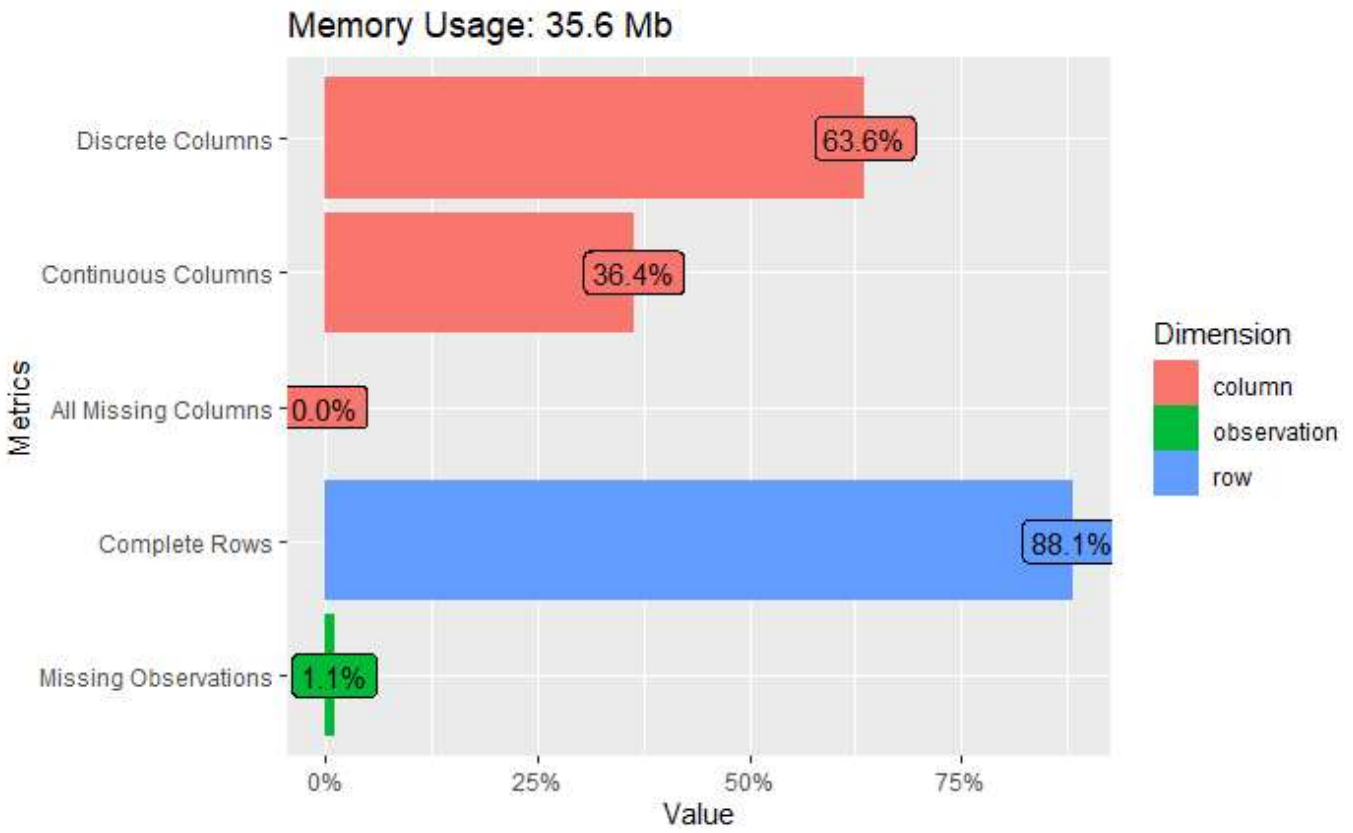
```
plot_missing(train_tbl)
```



Missing values in Credit Product column

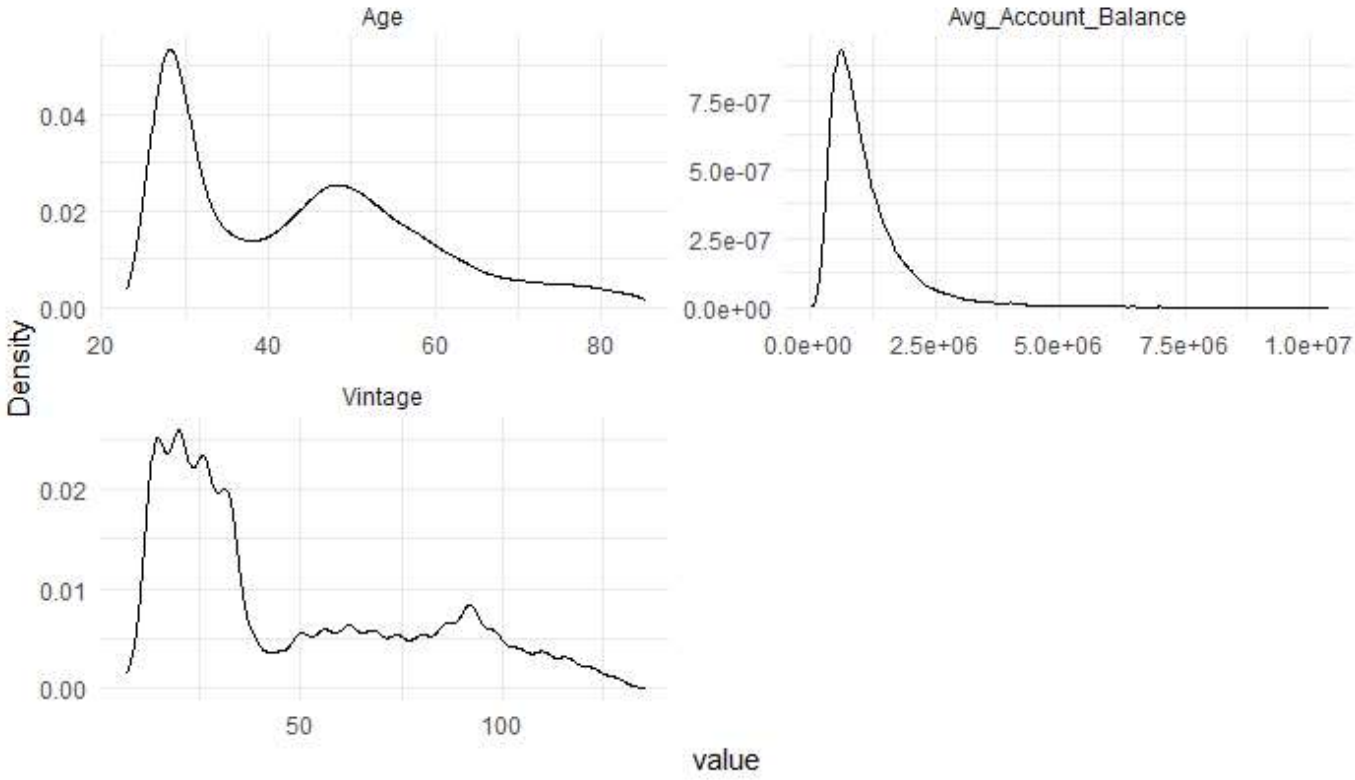
Hide

```
train_tbl %>%
  plot_intro()
```



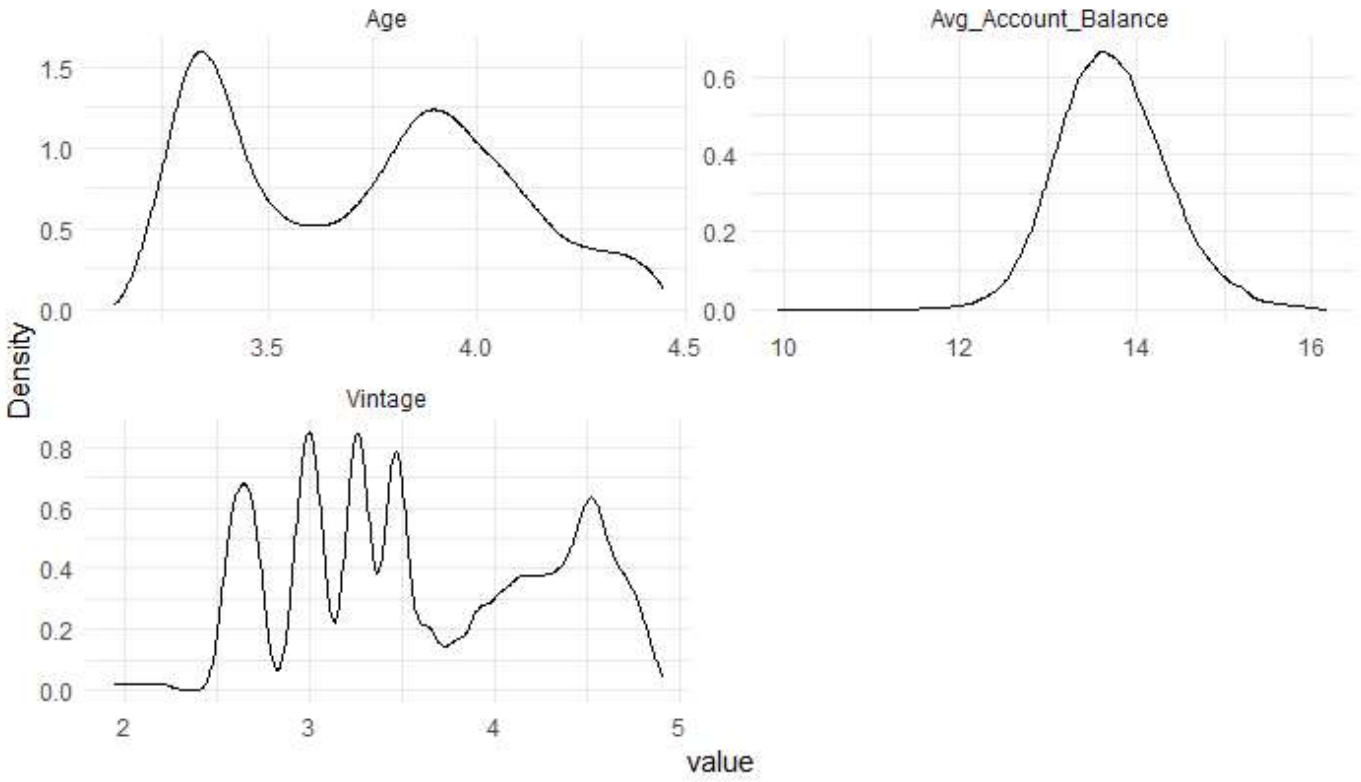
Hide

```
train_tbl %>%
#  mutate_if(is.numeric, log) %>%
  plot_density(ggtheme = theme_minimal(), ncol = 2)
```



Hide

```
train_tbl %>%
  mutate_if(is.numeric, log) %>%
  plot_density(ggtheme = theme_minimal(), ncol = 2)
```

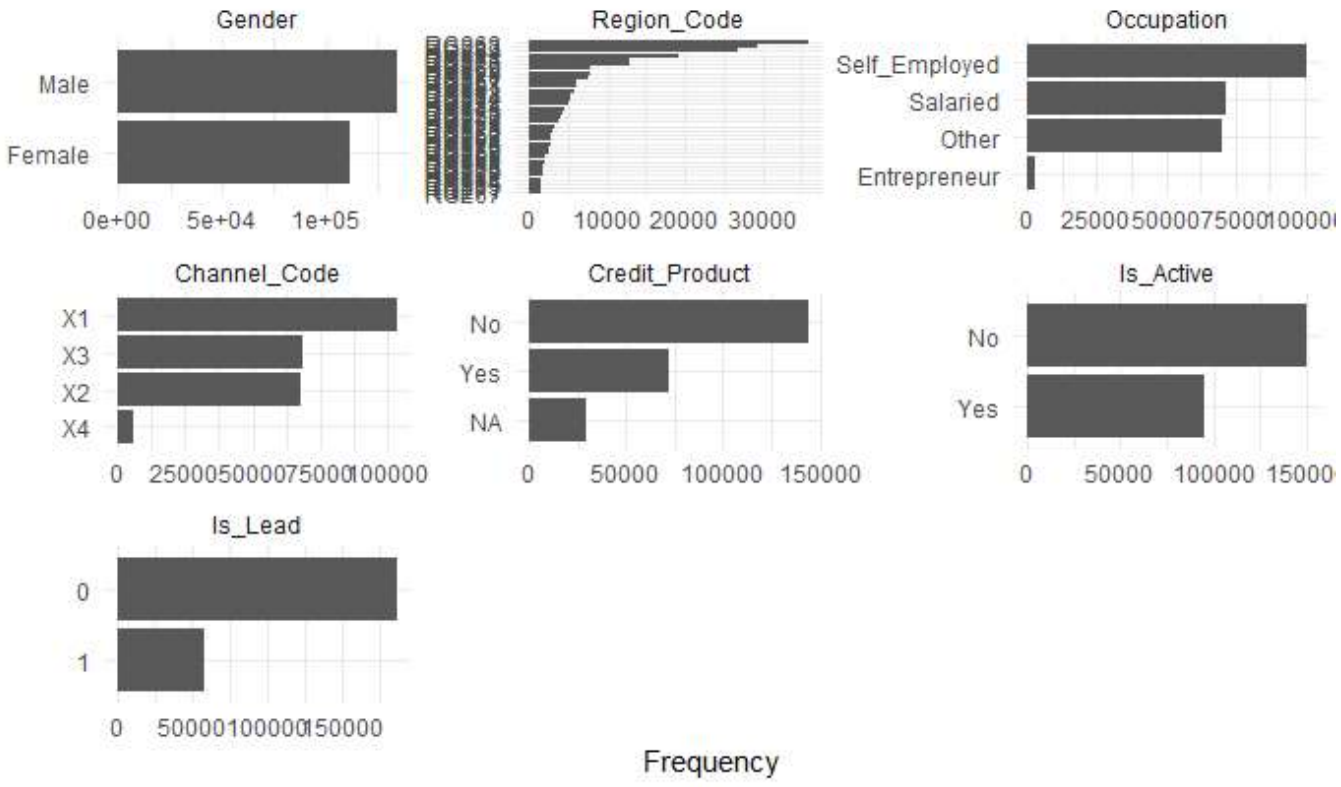


We can see that Age has two distinct distributions, hence we might need to discretize it. Also Avg Account Balance is skewed. Vintage has five distinct distributions.

Hide

```
train_tbl %>%
  plot_bar(ggtheme = theme_minimal())
```

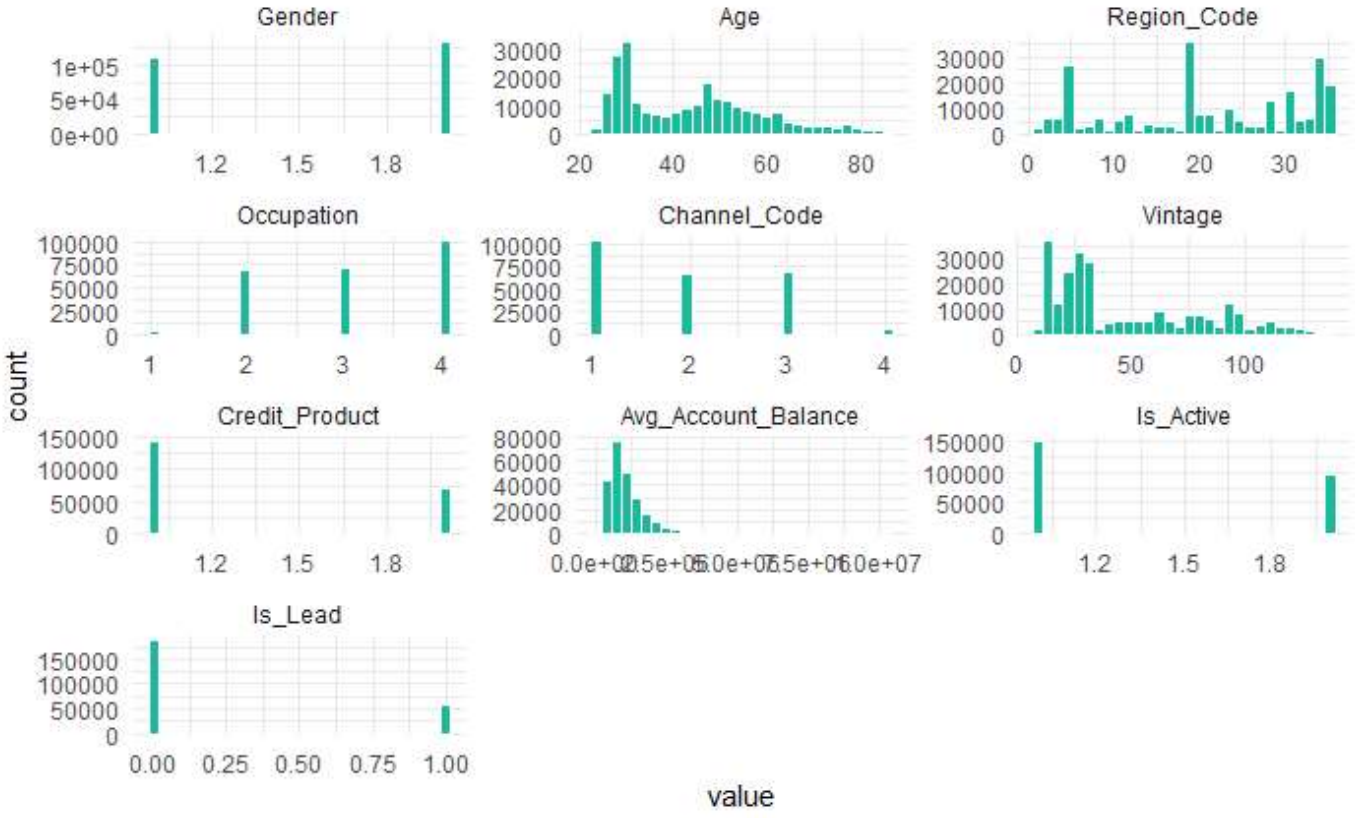
```
1 columns ignored with more than 50 categories.
ID: 245725 categories
```



We can see that there is an imbalance of classes

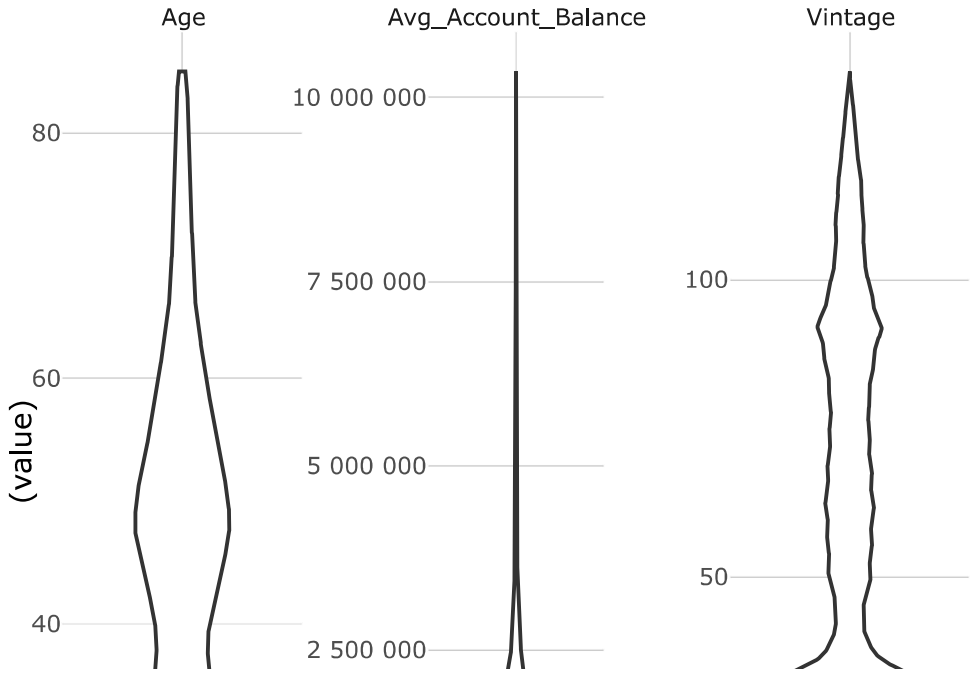
Hide

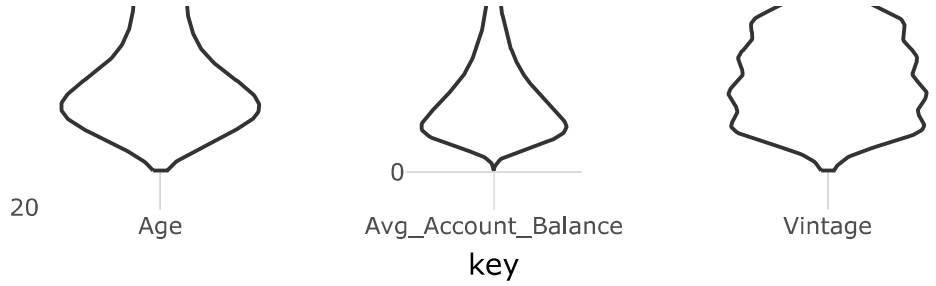
```
train_tbl %>%
  mutate_if(is.character, as.factor) %>%
  mutate_if(is.factor, as.numeric) %>%
  select(-ID) %>%
  gather("name", "value", factor_key = TRUE) %>%
  #  pivot_longer(cols = Gender:Is_Lead, names_to = "name", values_to = "value") %>%
  ggplot(aes(x = value, group = name)) +
  geom_histogram(bins = 30, fill = palette_light()[[3]], color = "white") +
  facet_wrap(~ name, ncol = 3, scale = "free") +
  theme_minimal()
```



Hide

```
ggplotly(
  train_tbl %>%
  select_if(is.numeric) %>%
  select(-Is_Lead) %>%
  gather() %>%
  ggplot(mapping = aes(key, (value))) +
  geom_violin(draw_quantiles = TRUE) +
  facet_wrap(~ key, scales = "free") +
  scale_y_continuous(labels = scales::number_format()) +
  theme_minimal())
```





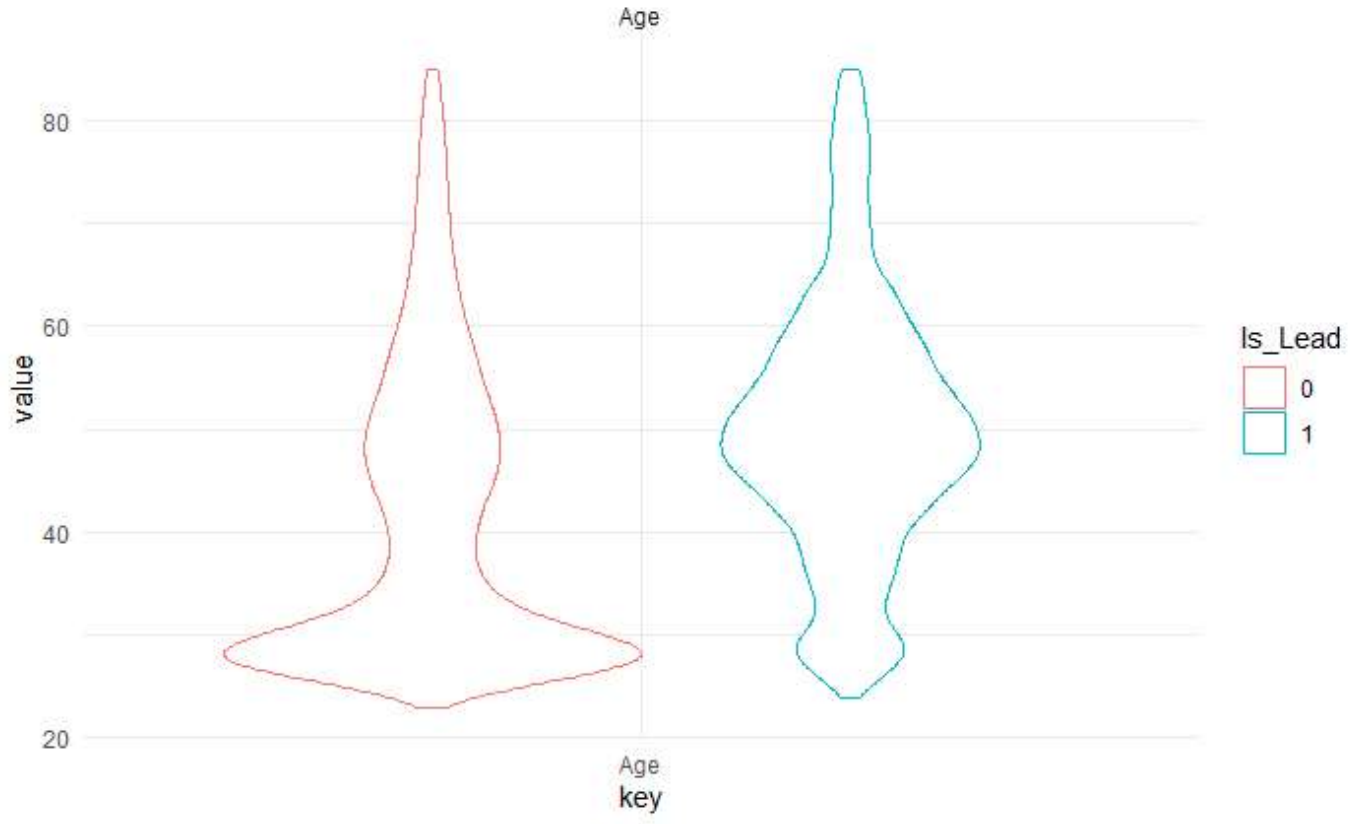
Hide

```
NA
```

Again, Age has two distinct distributions. Also Avg Account Balance is skewed. Vintage has five distinct distributions.

Hide

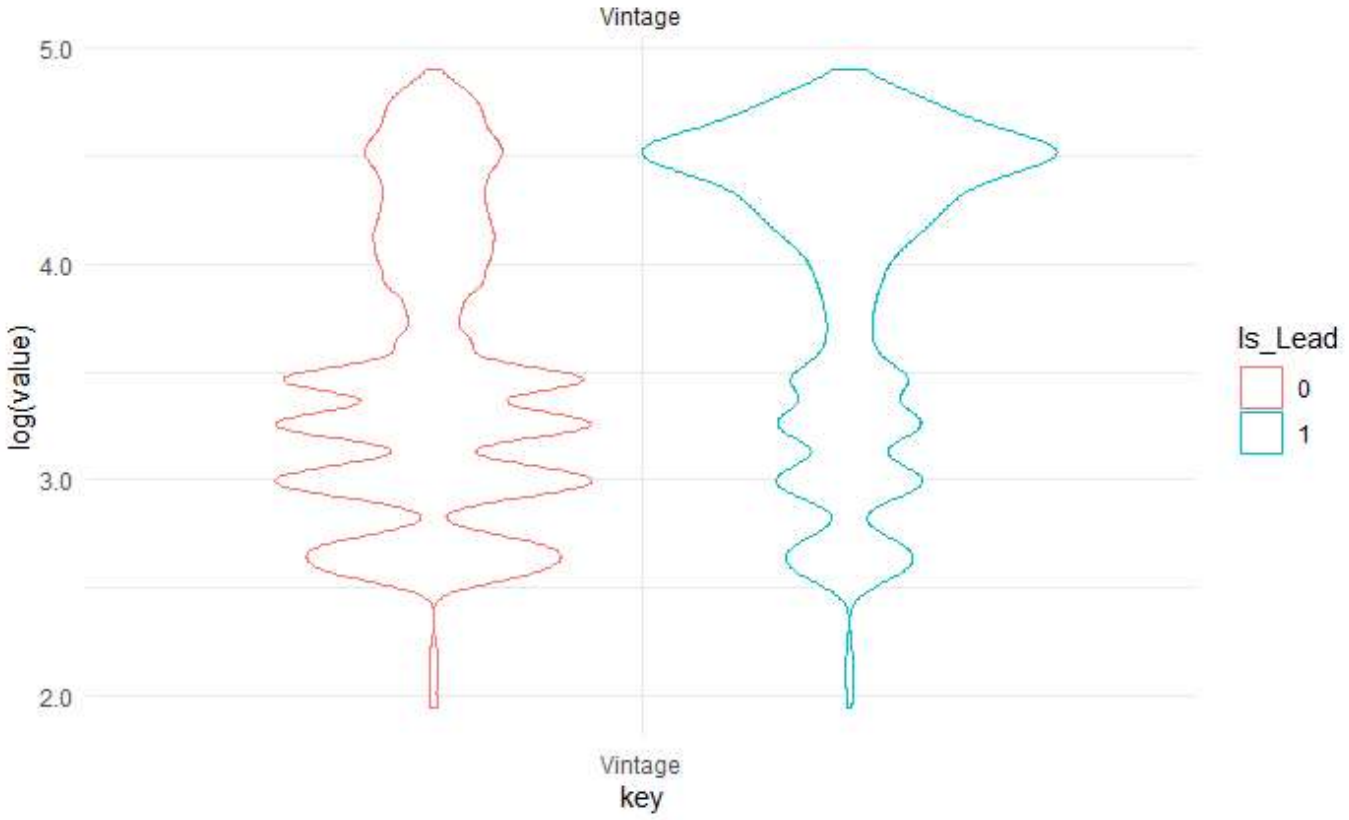
```
train_tbl %>%
  mutate(Is_Lead = as.factor(Is_Lead)) %>%
  select(Is_Lead, Age) %>%
  gather("key", "value", 2:ncol(.)) %>%
  ggplot(mapping = aes(key, value, color = Is_Lead)) +
  geom_violin(draw_quantiles = TRUE) +
  facet_wrap(~ key, scales = "free") +
  scale_y_continuous(labels = scales::number_format()) +
  theme_minimal()
```



We now know that leads are of higher age (between 40 to 60)

Hide

```
train_tbl %>%
  mutate(Is_Lead = as.factor(Is_Lead)) %>%
  select(Is_Lead, Vintage) %>%
  gather("key", "value", 2:ncol(.)) %>%
  ggplot(mapping = aes(key, log(value), color = Is_Lead)) +
  geom_violin(draw_quantiles = TRUE) +
  facet_wrap(~ key, scales = "free", ncol = 2) +
  scale_y_continuous(labels = scales::number_format()) +
  theme_minimal()
```

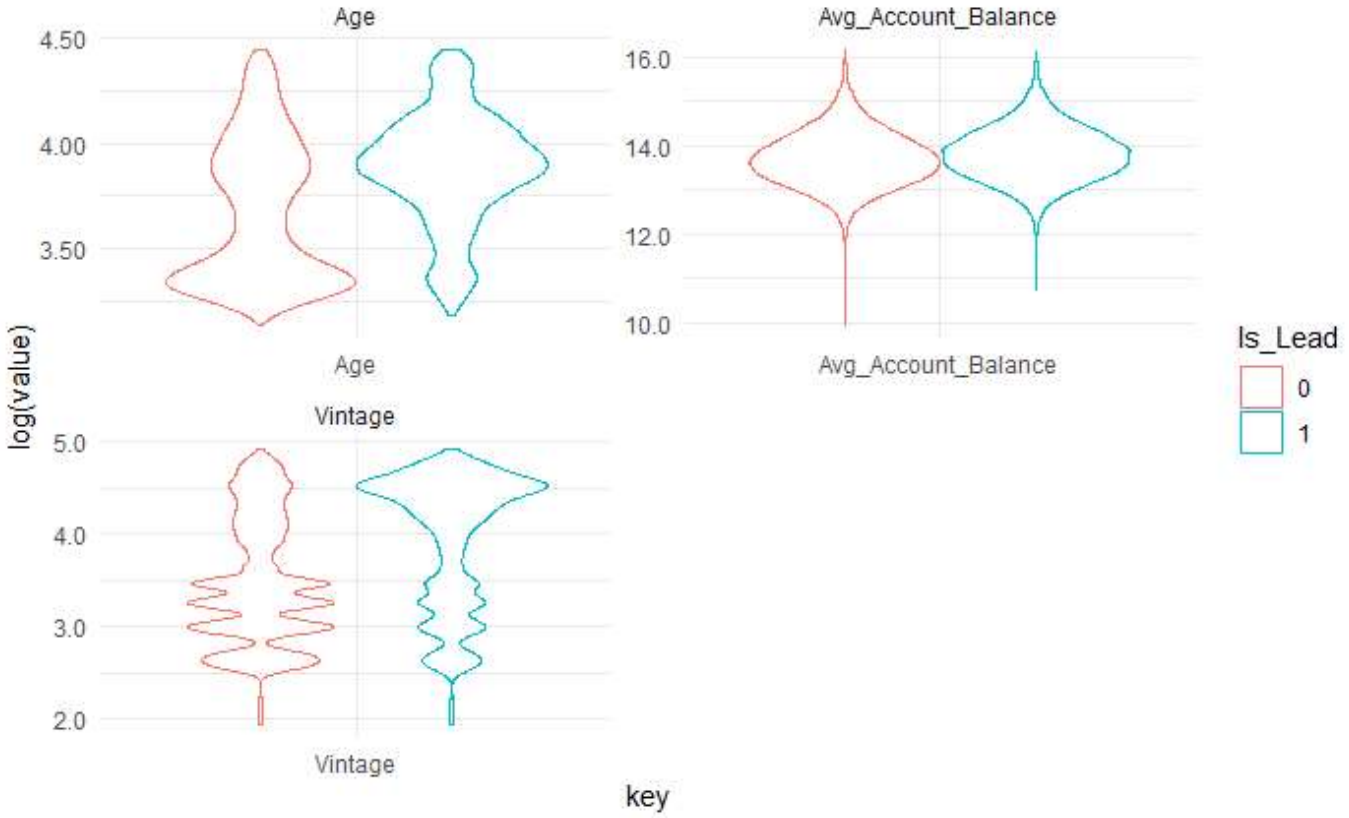


Hide

```
NA
NA
```

Hide

```
train_tbl %>%
  mutate(Is_Lead = as.factor(Is_Lead)) %>%
  select(Is_Lead, Age, Avg_Account_Balance, Vintage) %>%
  gather("key", "value", 2:4) %>%
  ggplot(mapping = aes(key, log(value), color = Is_Lead)) +
  geom_violin(draw_quantiles = TRUE) +
  facet_wrap(~ key, scales = "free", ncol = 2) +
  scale_y_continuous(labels = scales::number_format()) +
  theme_minimal()
```



Some kind of transform like Log can be of use in binning features Age and Vintage

Hide

```
train_tbl %>%
  count(Is_Lead) %>%
  mutate(pct = n/sum(n))
```

Is_Lead <dbl>	n <int>	pct <dbl>
0	187437	0.7627917
1	58288	0.2372083

2 rows

Hide

```
NA
```

There is imbalance of target variable

Hide

```
train_tbl %>%
  select_if(is.character) %>%
  select(-ID) %>%
  map(~ table(.) %>% prop.table(.))
```

```
$Gender
.
   Female      Male
0.4538732 0.5461268

$Region_Code
.
      RG250       RG251       RG252       RG253       RG254       RG255       RG256       RG2
57       RG258       RG259       RG260       RG261
0.010157697 0.024214060 0.017442263 0.007561298 0.109227795 0.008212433 0.011586123 0.0248285
69 0.007939770 0.010523960 0.012656425 0.031063180
      RG262       RG263       RG264       RG265       RG266       RG267       RG268       RG2
69       RG270       RG271       RG272       RG273
0.007276427 0.015004578 0.011366365 0.006291586 0.006421813 0.006092176 0.146236647 0.0319991
86 0.031417235 0.006275308 0.021373487 0.018300946
      RG274       RG275       RG276       RG277       RG278       RG279       RG280       RG2
81       RG282       RG283       RG284
0.021511853 0.013205820 0.011248347 0.052196561 0.007414793 0.016180690 0.051989012 0.0207264
22 0.023721640 0.119711059 0.078624479

$Occupation
.
 Entrepreneur          Other      Salaried Self_Employed
    0.0108536     0.2855753     0.2930064     0.4105647

$Channel_Code
.
        X1         X2         X3         X4
0.42208973 0.27561705 0.27962967 0.02266355

$Credit_Product
.
       No       Yes
0.6670841 0.3329159

$Is_Active
.
       No       Yes
0.6116187 0.3883813
```

Hide

```
train_tbl %>%
  group_by(Occupation, Is_Lead) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  group_by(Occupation) %>%
  mutate(pct = n/sum(n))
```

```
`summarise()` has grouped output by 'Occupation'. You can override using the `.groups` argume
nt.
```

Occupation	Is_Lead	n	pct
<chr>	<dbl>	<int>	<dbl>

Occupation <chr>	Is_Lead <dbl>	n <int>	pct <dbl>
Entrepreneur	0	905	0.3393326
Entrepreneur	1	1762	0.6606674
Other	0	52984	0.7550482
Other	1	17189	0.2449518
Salaried	0	60503	0.8403311
Salaried	1	11496	0.1596689
Self_Employed	0	73045	0.7240350
Self_Employed	1	27841	0.2759650

8 rows

Enterpreneur are going to be more likely to be a lead than other professions

Hide

```
train_tbl %>%
  group_by(Credit_Product, Is_Lead) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  group_by(Credit_Product) %>%
  mutate(pct = n/sum(n))
```

```
`summarise()` has grouped output by 'Credit_Product'. You can override using the `.groups` ar
gument.
```

Credit_Product <chr>	Is_Lead <dbl>	n <int>	pct <dbl>
No	0	133734	0.9264116
No	1	10623	0.0735884
Yes	0	49353	0.6850492
Yes	1	22690	0.3149508
NA	0	4350	0.1483376
NA	1	24975	0.8516624

6 rows

Lot of missing Credit Products are leads … hence, we need to impute this missing value somehow.

Hide

```
train_tbl %>%
  group_by(Channel_Code, Is_Lead) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  group_by(Channel_Code) %>%
  mutate(pct = n/sum(n))
```

```
`summarise()` has grouped output by 'Channel_Code'. You can override using the `.groups` argu
ment.
```

Channel_Code <chr>	Is_Lead <dbl>	n <int>	pct <dbl>
X1	0	94236	0.90857903
X1	1	9482	0.09142097
X2	0	45519	0.67210525
X2	1	22207	0.32789475
X3	0	43493	0.63297532
X3	1	25219	0.36702468
X4	0	4189	0.75219968
X4	1	1380	0.24780032

8 rows

X1 is a bad channel to find a lead

Trimming of useless things pre - data Modeling :

Hide

```
train_tbl <- train_tbl %>%
  select(-ID) %>%
  mutate(Gender = as.factor(Gender),
         Region_Code = as.factor(Region_Code),
         Occupation = as.factor(Occupation),
         Channel_Code = as.factor(Channel_Code),
         Credit_Product = as.factor(Credit_Product),
         Is_Active = as.factor(Is_Active),
         Is_Lead = as.factor(Is_Lead))

test_tbl <- test_tbl %>%
  select(-ID) %>%
  mutate(Gender = as.factor(Gender),
         Region_Code = as.factor(Region_Code),
         Occupation = as.factor(Occupation),
         Channel_Code = as.factor(Channel_Code),
         Credit_Product = as.factor(Credit_Product),
         Is_Active = as.factor(Is_Active))
```

Modeling with AutoML

Hide

```
print(h2o.auc(model_automl@leader, valid = TRUE))
```

```
[1] 0.8729794
```

Making Predictions

Hide

```
final_df <- h2o.predict(aml_leader, newdata = test_tbl) %>% as_tibble()
```

```
  |
  |
|   0%
  |
  |=======================================================================================
=======================================================================================
=========================================================| 100%
```

Hide

```
final_df <- h2o.predict(aml_leader, newdata = test_tbl) %>% as_tibble()
```

```
  |
  |
|   0%
  |
  |=======================================================================================
=======================================================================================
=============================================================| 100%
```

Hide

```
final_df
```

predict <fctr>	p0 <dbl>	p1 <dbl>
0	0.95728920	0.04271080
1	0.16559914	0.83440086
0	0.95429798	0.04570202
0	0.97258683	0.02741317
0	0.98076416	0.01923584
0	0.92447608	0.07552392
0	0.93752274	0.06247726
0	0.95601912	0.04398088
1	0.04824041	0.95175959
0	0.82723941	0.17276059

1-10 of 105,312 rows Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
final_tbl <- bind_cols(read_csv("../data/test_mSzZ8RL.csv"), final_df %>% select(p1))
```

```
-- Column specification --------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------
cols(
  ID = col_character(),
  Gender = col_character(),
  Age = col_double(),
  Region_Code = col_character(),
  Occupation = col_character(),
  Channel_Code = col_character(),
  Vintage = col_double(),
  Credit_Product = col_character(),
  Avg_Account_Balance = col_double(),
  Is_Active = col_character()
)
```

Hide

```
final_tbl <- bind_cols(read_csv("../data/test_mSzZ8RL.csv"), final_df %>% select(p1))
```

```
-- Column specification --------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------
cols(
  ID = col_character(),
  Gender = col_character(),
  Age = col_double(),
  Region_Code = col_character(),
  Occupation = col_character(),
  Channel_Code = col_character(),
  Vintage = col_double(),
  Credit_Product = col_character(),
  Avg_Account_Balance = col_double(),
  Is_Active = col_character()
)
```

Hide

```
submission_file <- final_tbl %>%
  rename(Is_Lead = p1) %>%
  #  mutate(Is_Lead = ifelse(Is_Lead == "Yes", 1, 0)) %>%
  select(ID, Is_Lead)
final_tbl <- bind_cols(read_csv("../data/test_mSzZ8RL.csv"), final_df %>% select(p1))
```

```
-- Column specification --------------------------------------------------------
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
cols(
  ID = col_character(),
  Gender = col_character(),
  Age = col_double(),
  Region_Code = col_character(),
  Occupation = col_character(),
  Channel_Code = col_character(),
  Vintage = col_double(),
  Credit_Product = col_character(),
  Avg_Account_Balance = col_double(),
  Is_Active = col_character()
)
```

Hide

```
submission_file <- final_tbl %>%
  rename(Is_Lead = p1) %>%
  #  mutate(Is_Lead = ifelse(Is_Lead == "Yes", 1, 0)) %>%
  select(ID, Is_Lead)
write_csv(submission_file, "../output/submission_h2o_aml_scenario120.csv")
```