

# DATA 621 - HW4

Andrew Bowen, Glen Davis, Shoshana Farber, Joshua Forster, Charles Ugiagbe

2023-10-30

## Homework 4 - Binary Logistic Regression & Multiple Linear Regression

### Data Exploration:

We load an auto insurance company dataset containing 8,161 records. Each record represents a customer, and each record has two response variables: **TARGET\_FLAG** and **TARGET\_AMT**. Below is a short description of all the variables of interest in the data set, including these response variables:

VARIABLE NAME	DEFINITION
INDEX	Identification Variable
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO
TARGET_AMT	If car was in a crash, what was the cost
AGE	Age of Driver
BLUEBOOK	Value of Vehicle
CAR_AGE	Vehicle Age
CAR_TYPE	Type of Car
CAR_USE	Vehicle Use
CLM_FREQ	# Claims (Past 5 Years)
EDUCATION	Max Education Level
HOMEKIDS	# Children at Home
HOME_VAL	Home Value
INCOME	Income
JOB	Job Category
KIDSDRIV	# Driving Children
MSTATUS	Marital Status
MVR_PTS	Motor Vehicle Record Points
OLDCLAIM	Total Claims (Past 5 Years)
PARENT1	Single Parent
RED_CAR	A Red Car
REVOKED	License Revoked (Past 7 Years)
SEX	Gender
TIF	Time in Force
TRAVTIME	Distance to Work
URBANICITY	Home/Work Area
YOJ	Years on Job

We take a look at the classes of our variables.

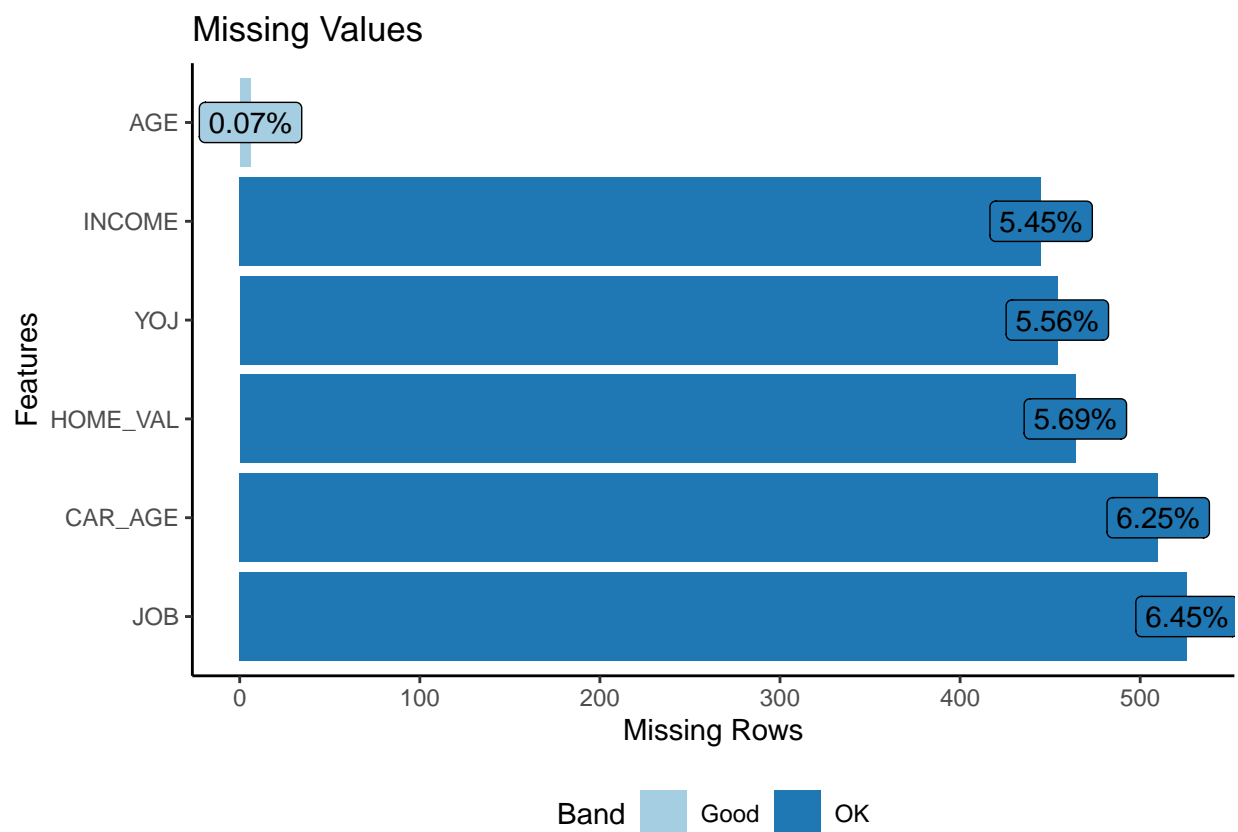
Class	Count	Variables
character	14	BLUEBOOK, CAR_TYPE, CAR_USE, EDUCATION, HOME_VAL, INCOME, JOB, MSTATUS, O
integer	11	AGE, CAR_AGE, CLM_FREQ, HOMEKIDS, INDEX, KIDSDRIV, MVR_PTS, TARGET_FLAG, T
numeric	1	TARGET_AMT

INCOME, HOME\_VAL, BLUEBOOK, and OLDCLAIM are all character columns that need to be recoded as integers. TARGET\_FLAG and the remaining character columns will all need to be recoded as factors.

We remove the identification variable INDEX and take a look at a summary of the dataset's completeness.

rows	8161
columns	25
all_missing_columns	0
total_missing_values	2405
complete_rows	6045

None of our columns are completely devoid of data. There are 6,045 complete rows in the dataset, which is about 74% of our observations. There are 2,405 total missing values. We take a look at which variables contain these missing values and what the spread is.



A very small percentage of observations contain missing AGE values. The INCOME, YOJ, HOME\_VAL, CAR\_AGE, and JOB variables are each missing around 5.5 to 6.5 percent of values. There are no variables containing such extreme proportions of missing values that removal would be warranted on that basis alone.

We have 14 numerical variables and 11 categorical variables (including the dummy variable **TARGET\_FLAG**). We recode the categorical variables as factors and list the possible ranges or values for each variable in the breakdown below:

Variable	Type	Values
AGE	Numeric	16 - 81
BLUEBOOK	Numeric	1500 - 69740
CAR_AGE	Numeric	-3 - 28
CLM_FREQ	Numeric	0 - 5
HOME_VAL	Numeric	0 - 885282
HOMEKIDS	Numeric	0 - 5
INCOME	Numeric	0 - 367030
KIDSDRIV	Numeric	0 - 4
MVR_PTS	Numeric	0 - 13
OLDCLAIM	Numeric	0 - 57037
TARGET_AMT	Numeric	0 - 107586.1
TIF	Numeric	1 - 25
TRAVTIME	Numeric	5 - 142
YOJ	Numeric	0 - 23
CAR_TYPE	Categorical	Minivan, Panel Truck, Pickup, Sports Car, Van, z_SUV
CAR_USE	Categorical	Commercial, Private
EDUCATION	Categorical	<High School, Bachelors, Masters, PhD, z_High School
JOB	Categorical	Clerical, Doctor, Home Maker, Lawyer, Manager, Professional, Student, z_Blue Collar
MSTATUS	Categorical	Yes, z_No
PARENT1	Categorical	No, Yes
RED_CAR	Categorical	no, yes
REVOKED	Categorical	No, Yes
SEX	Categorical	M, z_F
TARGET_FLAG	Categorical	0, 1
URBANICITY	Categorical	Highly Urban/ Urban, z_Highly Rural/ Rural

**TK: Some of the factor levels are named inconsistently, so we rename those.**

Let's take a look at the summary statistics for each variable.

```
## TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS
## 0:6008 Min. : 0 Min. :0.0000 Min. :16.00 Min. :0.0000
## 1:2153 1st Qu.: 0 1st Qu.:0.0000 1st Qu.:39.00 1st Qu.:0.0000
## Median : 0 Median :0.0000 Median :45.00 Median :0.0000
## Mean : 1504 Mean :0.1711 Mean :44.79 Mean :0.7212
## 3rd Qu.: 1036 3rd Qu.:0.0000 3rd Qu.:51.00 3rd Qu.:1.0000
## Max. :107586 Max. :4.0000 Max. :81.00 Max. :5.0000
## NA's :6
## YOJ INCOME PARENT1 HOME_VAL MSTATUS
## Min. : 0.0 Min. : 0 No :7084 Min. : 0 Yes :4894
## 1st Qu.: 9.0 1st Qu.: 28097 Yes:1077 1st Qu.: 0 z_No:3267
## Median :11.0 Median : 54028 Median :161160
## Mean :10.5 Mean : 61898 Mean :154867
## 3rd Qu.:13.0 3rd Qu.: 85986 3rd Qu.:238724
## Max. :23.0 Max. :367030 Max. :885282
## NA's :454 NA's :445 NA's :464
## SEX EDUCATION JOB TRAVTIME
## M :3786 <High School :1203 z_Blue Collar:1825 Min. : 5.00
```

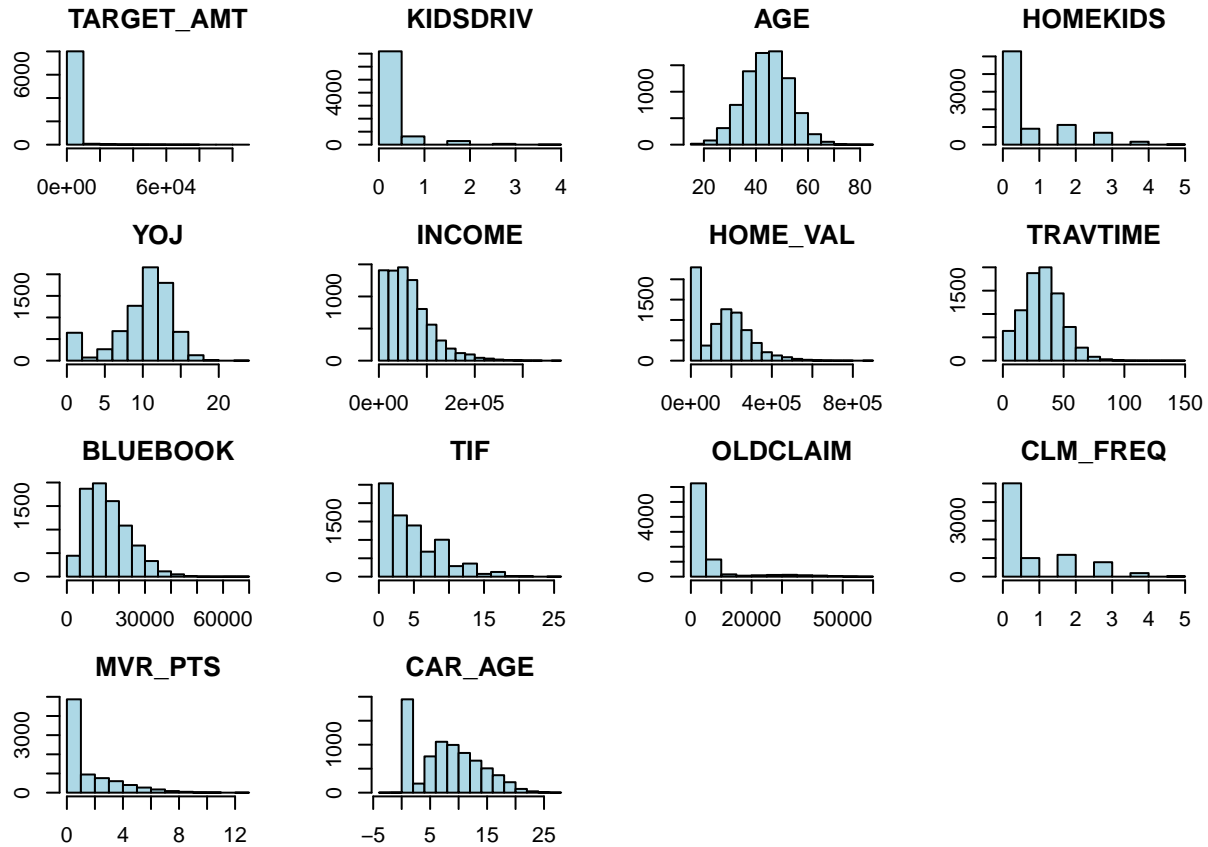
```

## z_F:4375 Bachelors :2242 Clerical :1271 1st Qu.: 22.00
## Masters :1658 Professional :1117 Median : 33.00
## PhD : 728 Manager : 988 Mean : 33.49
## z_High School:2330 Lawyer : 835 3rd Qu.: 44.00
## (Other) :1599 Max. :142.00
## NA's : 526
## CAR_USE BLUEBOOK TIF CAR_TYPE
## Commercial:3029 Min. : 1500 Min. : 1.000 Minivan :2145
## Private :5132 1st Qu.: 9280 1st Qu.: 1.000 Panel Truck: 676
## Median :14440 Median : 4.000 Pickup :1389
## Mean :15710 Mean : 5.351 Sports Car : 907
## 3rd Qu.:20850 3rd Qu.: 7.000 Van : 750
## Max. :69740 Max. :25.000 z_SUV :2294
##
## RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS
## no :5783 Min. : 0 Min. :0.0000 No :7161 Min. : 0.000
## yes:2378 1st Qu.: 0 1st Qu.:0.0000 Yes:1000 1st Qu.: 0.000
## Median : 0 Median :0.0000 Median : 1.000
## Mean : 4037 Mean :0.7986 Mean : 1.696
## 3rd Qu.: 4636 3rd Qu.:2.0000 3rd Qu.: 3.000
## Max. :57037 Max. :5.0000 Max. :13.000
##
## CAR_AGE URBANICITY
## Min. :-3.000 Highly Urban/ Urban :6492
## 1st Qu.: 1.000 z_Highly Rural/ Rural:1669
## Median : 8.000
## Mean : 8.328
## 3rd Qu.:12.000
## Max. :28.000
## NA's :510

```

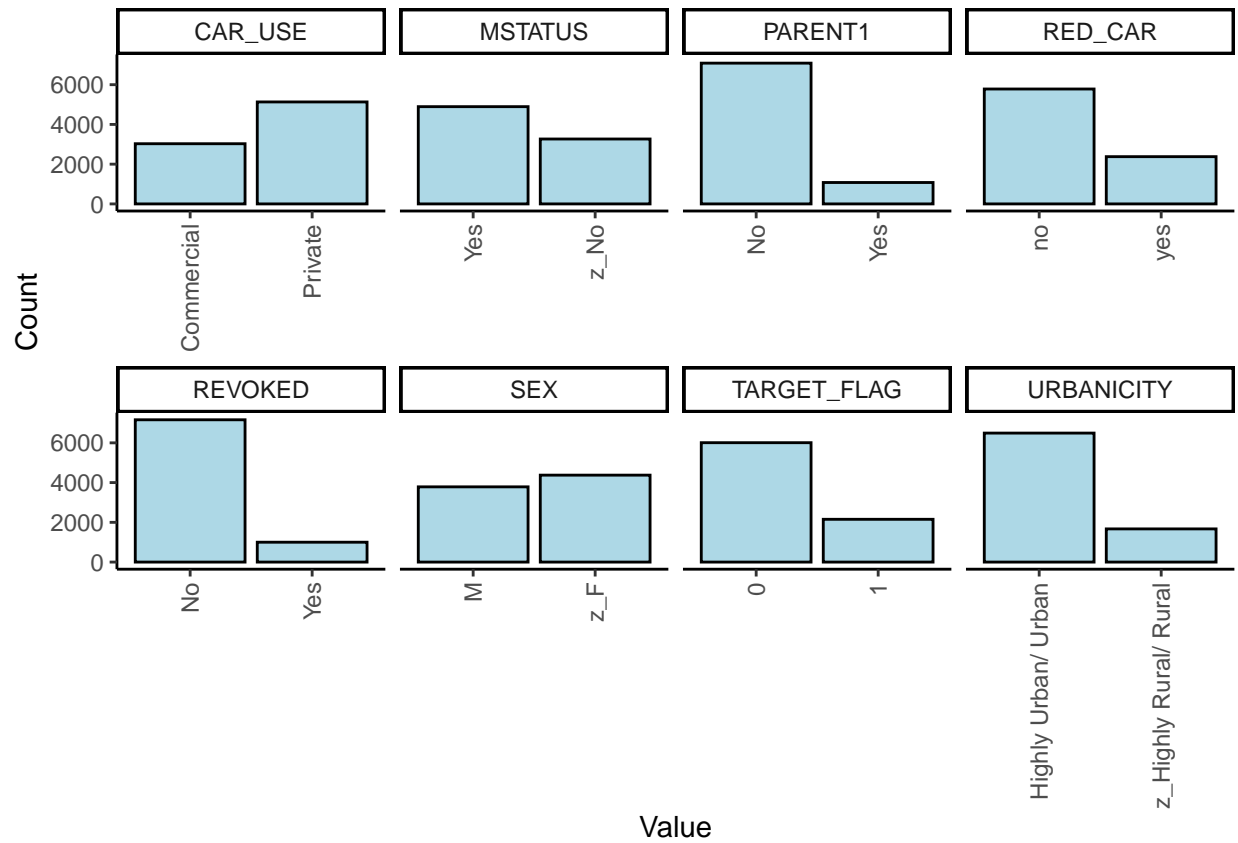
There are 6 NAs in AGE, 454 in YOJ, and 510 in CAR\_AGE.

Let's take a look at the distributions of the numeric variables.

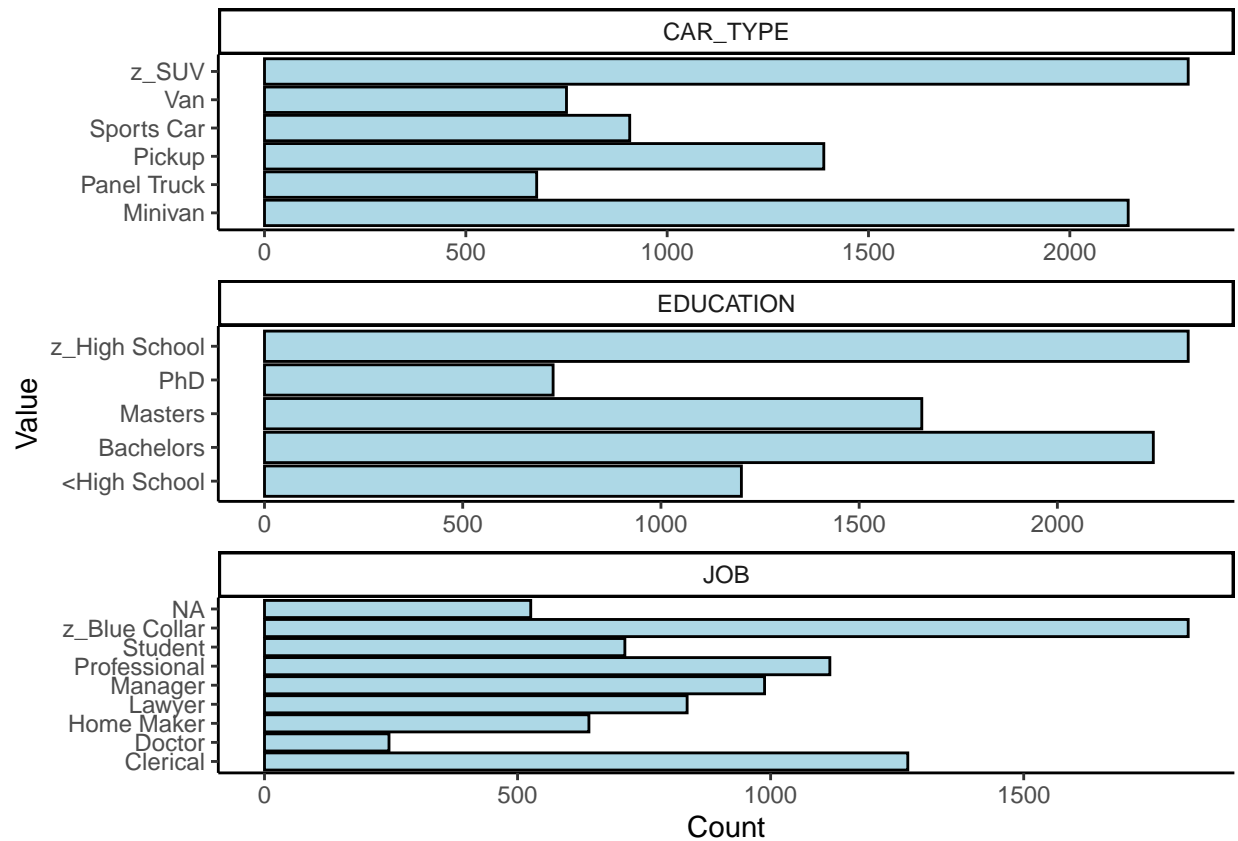


Let's also take a look at the distributions of the categorical variables. First, we look at the distributions for categorical variables with only two levels.

```
## 'summarise()' has grouped output by 'Variable'. You can override using the
## '.groups' argument.
```



Next we look at the distributions for the categorical variables with more than two levels.



## Data Preparation

### TK: MICE Imputation

### Appendix: Report Code

Below is the code for this report to generate the models and charts above.

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(DataExplorer)
library(knitr)

cur_theme <- theme_set(theme_classic())

my_url <- "https://raw.githubusercontent.com/andrewbowen19/businessAnalyticsDataMiningDATA621/main/data"
df <- read.csv(my_url, na.strings = "")

classes <- as.data.frame(unlist(lapply(df, class))) |>
  rownames_to_column()
cols <- c("Variable", "Class")
colnames(classes) <- cols
classes_summary <- classes |>
  group_by(Class) |>
```

```

    summarize(Count = n(),
              Variables = paste(sort(unique(Variable)), collapse = ", "))
knitr::kable(classes_summary, format = "simple")

vars <- c("INCOME", "HOME_VAL", "BLUEBOOK", "OLDCLAIM")
df <- df |>
  mutate(across(all_of(vars), ~gsub("\\$|", "", .) |> as.integer()))

train_df <- df |>
  select(-INDEX)
remove <- c("discrete_columns", "continuous_columns",
            "total_observations", "memory_usage")
completeness <- introduce(train_df) |>
  select(-all_of(remove))
knitr::kable(t(completeness), format = "simple")

p1 <- plot_missing(train_df, missing_only = TRUE,
                  ggtheme = theme_classic(), title = "Missing Values")

p1 <- p1 +
  scale_fill_brewer(palette = "Paired")
p1

output <- split_columns(train_df, binary_as_factor = TRUE)
num <- data.frame(Variable = names(output$continuous),
                  Type = rep("Numeric", ncol(output$continuous)))
cat <- data.frame(Variable = names(output$discrete),
                  Type = rep("Categorical", ncol(output$discrete)))
ranges <- as.data.frame(t(sapply(train_df |> select(-names(output$discrete)),
                                range, na.rm = TRUE)))
factors <- names(output$discrete)
train_df <- train_df |>
  mutate(across(all_of(factors), ~as.factor(.)))
values <- as.data.frame(t(sapply(train_df |> select(all_of(factors)),
                                levels)))

values <- values |>
  mutate(across(all_of(factors), ~toString(unlist(.))))
values <- as.data.frame(t(values)) |>
  rownames_to_column()
cols <- c("Variable", "Values")
colnames(values) <- cols
remove <- c("V1", "V2")
ranges <- ranges |>
  rownames_to_column() |>
  group_by(rowname) |>
  mutate(Values = toString(c(V1, " - ", round(V2, 1))),
         Values = str_replace_all(Values, ",", "")) |>
  select(-all_of(remove))
colnames(ranges) <- cols
num <- num |>
  merge(ranges)
cat <- cat |>
  merge(values)

```



```

num_vs_cat <- num |>
  bind_rows(cat)
knitr::kable(num_vs_cat, format = "simple")

summary(train_df)

# just numeric variables
numeric_train <- train_df[,sapply(train_df, is.numeric)]
par(mfrow=c(4,4))
par(mai=c(.3,.3,.3,.3))
variables <- names(numeric_train)
for (i in 1:(length(variables))) {
  hist(numeric_train[[variables[i]]], main = variables[i], col = "lightblue")
}

cat_pivot <- train_df |>
  select(all_of(factors)) |>
  pivot_longer(cols = all_of(factors),
               names_to = "Variable",
               values_to = "Value") |>
  group_by(Variable, Value) |>
  summarize(Count = n()) |>
  group_by(Variable) |>
  mutate(Levels = n()) |>
  ungroup()
p2 <- cat_pivot |>
  filter(Levels == 2) |>
  ggplot(aes(x = Value, y = Count)) +
  geom_col(fill = "lightblue", color = "black") +
  facet_wrap(vars(Variable), ncol = 4, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
p2

p3 <- cat_pivot |>
  filter(Levels > 2) |>
  ggplot(aes(x = Value, y = Count)) +
  geom_col(fill = "lightblue", color = "black") +
  coord_flip() +
  facet_wrap(vars(Variable), ncol = 1, scales = "free")
p3

```