Data Quality

# TASK 1 REPORT

# CONTENTS
# LIST OF IDEAS

**Artiom - Axe**

**Type of axe:**
- For bushcraft
- For chopping wood ● For splitting logs ● For firefighters.

Based on type we can specify some requirements and implicit and explicit expectations for it. I'm going to test classic chopping axe.

**Smoke:**
To keep it simple lets say that basic functionality of axe is to:
- Chop
- The handle has to be strong enough not to brake while using it (otherwise first step cannot be achieved)

**Critical path:**
- Durability of cutting age is extremely important. Therefore the brand of steel should be chosen wisely.
- Steel hardening process should be done in the proper way.

**Extended path:**
- Angle of the cutting edge should be selected based on the type of axe. Sharp and aggressive for chopping, wide for splitting, etc.
- The between cutting edge and handle should be calculated and applied correctly. It is a very important point in the long term because of pain and stress building up in hands after long usage. For example we need to chop down a couple of tries.
- Back side of the blade has to be heavy enough for us to increase our efficiency.
- Back side of the blade has to be in the golden spot for the chosen type, in order not to get stuck in wood or if the back side is needed for specific usage.
- Wedge should be inserted into the end of the handle to lock the head in place when assembling an axe. It goes into a slot, once the handle has been pressed into the axe head. This creates friction and locks the head in place.

# PYTHON DATA PROFILING

After creating the report in Command Prompt and opening it I can see anomalies in it.

**Age**

Minimum 5 values | Maximum 5 values

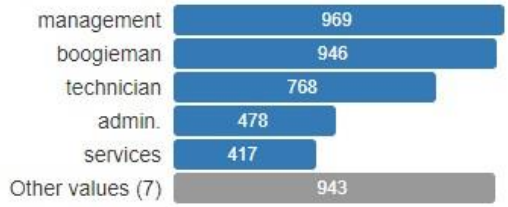
| Value | Count | Frequency (%) |
|---|---|---|
| -11 | 1 | < 0.1% |
| 3 | 1 | < 0.1% |
| 4 | 2 | < 0.1% |
| 6 | 1 | < 0.1% |
| 7 | 1 | < 0.1% |

Minimum 5 values | Maximum 5 values

| Value | Count | Frequency (%) |
|---|---|---|
| 200 | 1 | < 0.1% |
| 192 | 1 | < 0.1% |
| 160 | 1 | < 0.1% |
| 155 | 1 | < 0.1% |
| 150 | 1 | < 0.1% |

As we can see on the current screen there is certainly some broken data from the source. There are some limitations needed in source input.
Age -11 is completely out of this world. Therefore '-' shouldn't be allowed at all, only 10 numerals. **The proper ranges for numeric variables should be applied to all data.** In this case 18-100 or 21-100 years based on the law where the bank is located.

**JOB**




Because of big clusters of data I suppose that variants were generated by the bank and were offered to clients. Then I don't understand the necessity of creating the category 'boogieman'. **Categories should be better checked before going to production.**

**EDUCATION**

Same as in the last point. ‘Duck’ value should be considered as ‘Unknown’

## DEFAULT

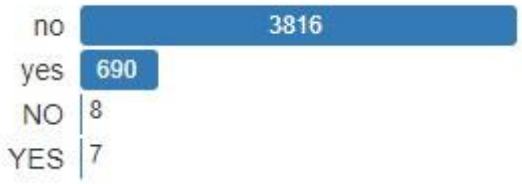

No abreviatures should be allowed here. I assume ‘N’ = ‘no’

## BALANCE

| | | | |
|---|---|---|---|
| **Distinct** | 2353 | **Minimum** | -3313 |
| **Distinct (%)** | 52.0% | **Maximum** | 71188 |
| **Missing** | 0 | **Zeros** | 357 |
| **Missing (%)** | 0.0% | **Zeros (%)** | 7.9% |
| **Infinite** | 0 | **Negative** | 366 |
| **Infinite (%)** | 0.0% | **Negative (%)** | 8.1% |
| **Mean** | 1422.657819 | **Memory size** | 35.4 KiB |

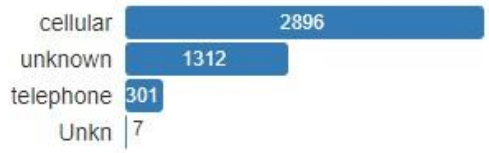
Negative balance is ok if we have credit and debit balance. I think whey should be marked with flags in this case. Also the amount of negative balances is lower than expected in this case.

## LOAN




**It’s better to unify possible answers.**

## CONTACT




Same here, ‘Unkn’ is a variant of ‘unknown’. The difference between ‘telephone’ and ‘cellular’ isn’t clear.

## DURATION

Should have the same covensions as ‘age’. Also measure of counting isn’t provided i cannot clearly say if ints in days or weeks(month and year would be impossible due values ile ‘3025’)

| | |
|---|---|
| **Minimum** | -999 |
| **Maximum** | 3025 |
| **Zeros** | 0 |
| **Zeros (%)** | 0.0% |
| **Negative** | 45 |
| **Negative (%)** | 1.0% |
| **Memory size** | 35.4 KiB |

**No missing cells constraint should be applied to all data.**

**All records with mistakes in them should be checked and corrected with customers individually or according to other reliable sources**.