**Flourishing in the Workplace – An exploration**

**Practice**

**Goal**

Apply the three steps of data analysis on the <Flourishing> dataset.

**Data exploration:**

The dataset is already clean and without NaN.

We can see that there are 11 columns.

We notice that "positivity" column is numeric with float but we can create a categorical column out of it : "positivity\_label".

Column “ID” will not help us as the number is irrelevant and not correlated to any other column.

**Exercise 1**

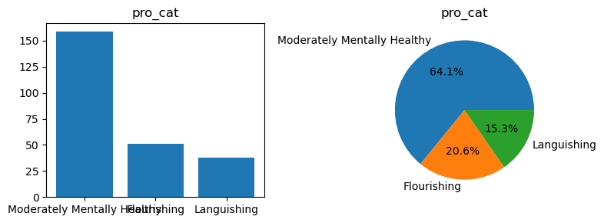
Explore univariate descriptive analysis.

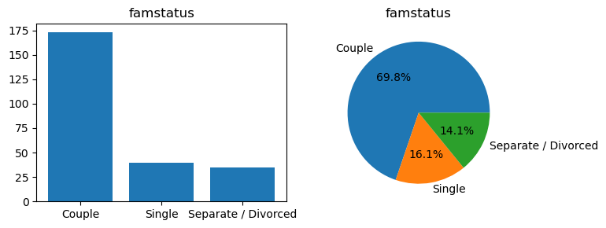
We can see on the plot that:

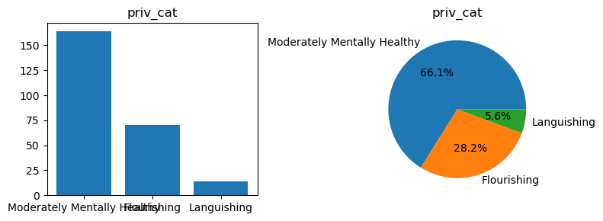
* Age:
  + Our population repartition looks like it follows a tendancy, with pick around 40
  + And noise around 50 and 55
* Education:
  + The majority of the sample Has completed 5 years after high school degree or more (= modality 6)
  + Modality 1 can be classed as outlier
* Sex:
  + This binary column is not balanced:
    - There are more female than male
* Famstatus:
  + 2 third of our population is Couple
  + 1/6 is separated/divorced
  + 1/6 is single
* Pro\_cat:
  + Roughly: population is separated in the same proportion as famstatus
* Pro\_quant:
  + Our population is more balanced than for “age”
  + But can see a pick around 46 and 66
* Priv\_cat:
  + 2/3 are moderate
  + Less than 1/3 are flourishing
  + And 5% only are languished
* Positivity: we can observe a very high peak around 2
* Flow: looks similar to priv\_quant

It seems that famstatus and pro\_cat could follow a same distribution.

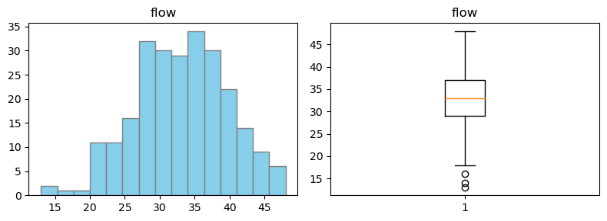
We may include priv\_cat in the same bag.

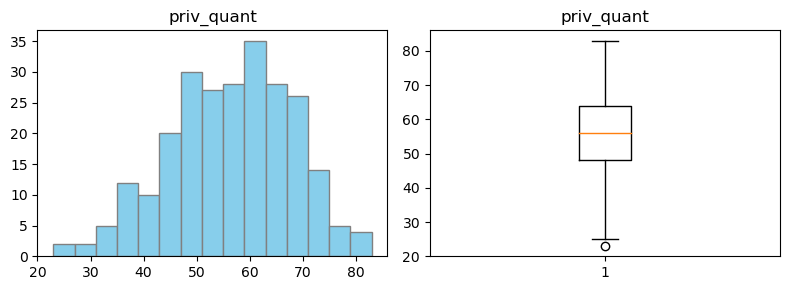






Priv\_quant and flow also looks similar:



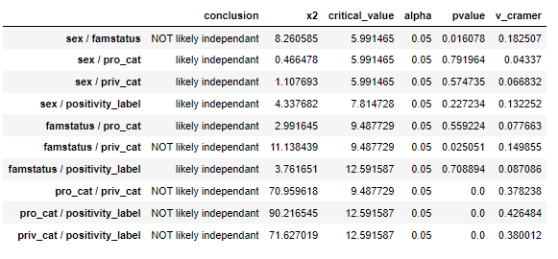


**Exercise 2**

Explore bivariate descriptive analysis.

1. If the variables are both categorical.
   1. Generate the contingency table
   2. Check the significance of their relationship with the chi-square test & provide Cramer’s V
   3. Plot their side-by-side bar charts
   4. Plot their stacked bar charts

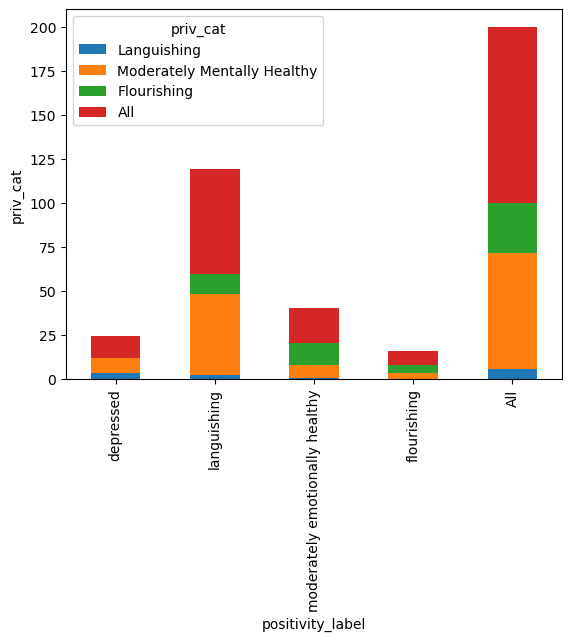
**The result of the analysis are :**

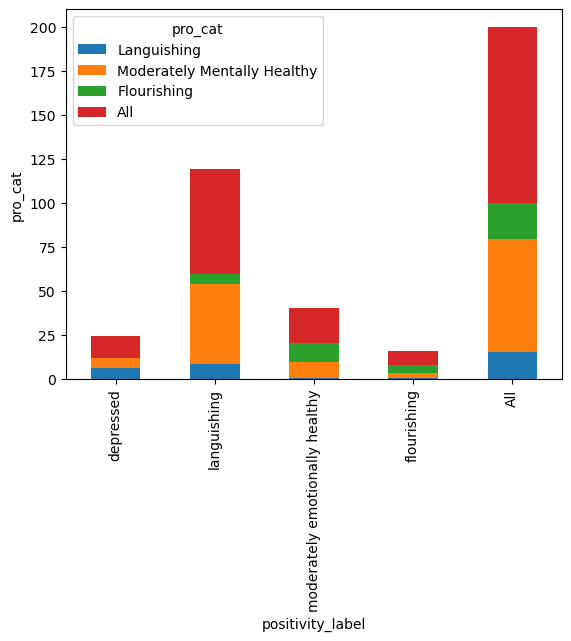


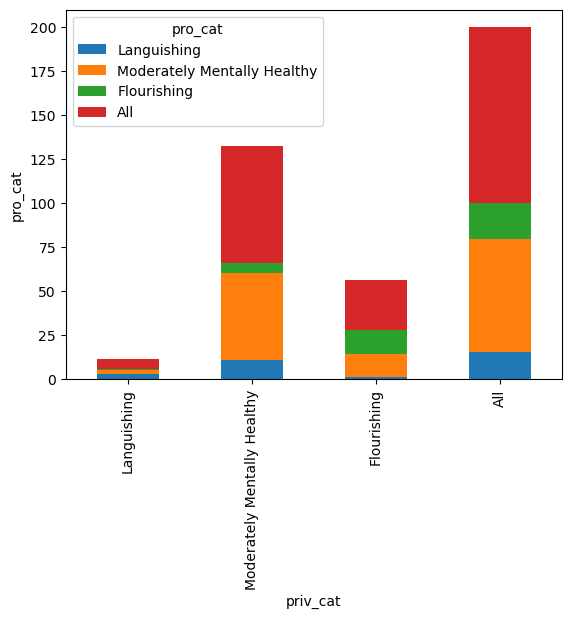
We can see that we haven’t been able to reject H0 for 5 couples. That means that tey are likely not independent.

But only the 3 last have a Cramer V > 0.3 which is medium.

Let’s have a look at their bar charts.

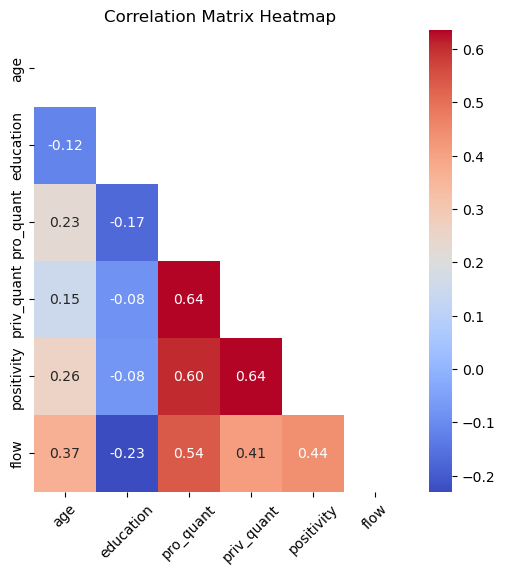






We can see that pro\_cat/positivity\_label and priv\_cat/positivity\_label looks similar and that pro\_cat/priv\_cat has the same shape.

1. If the variables are both quantitative.
   1. Compute the correlation (table)
   2. Check the significance of their relationship with the correlation test & provide the r value
   3. Plot their scatter plot (matrix)



The correlation matrix shows us that there we have interesting r especially for:

* Priv\_quant and pro\_quant
* Positivity and priv\_quant
* Positivity and pro\_quant
* Pro\_quant and flow
* Priv\_quant and flow
* Positivity and flow
* Age and flow

That’s much more relation than before.

Let’s run a T test to complete our study.



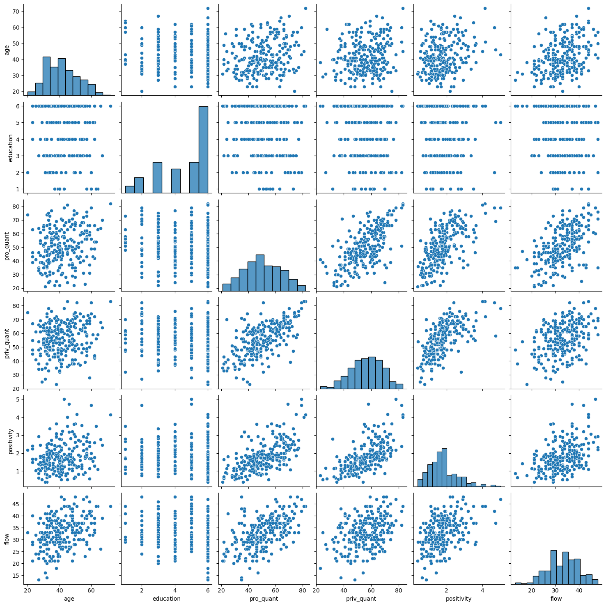
We can see that we cannot reject H0 for:

* Age/education
* Education / priv\_quant
* Education / positivity

This means they might have the same mean.

But also the r of those couple are not very high meaning that the relation is not that strong.

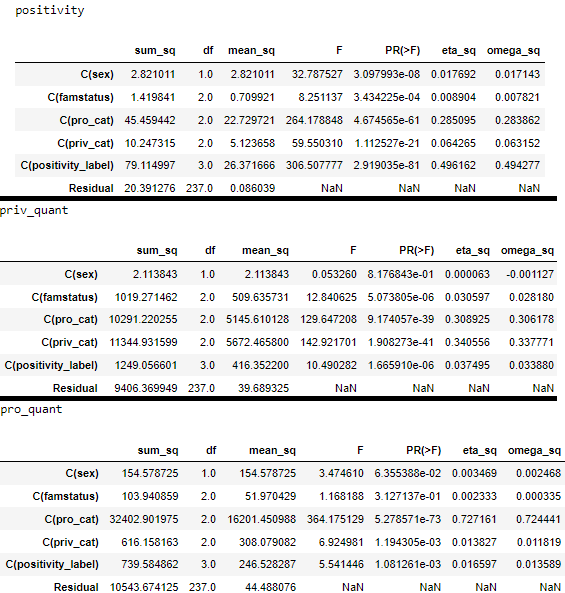
Let’s have a look at the scatter plot:



The scatter plot suggests that relation fond with the correlation matrix may be interesting for linear regression.

1. If the variables are mixed categorical & quantitative.
   1. Compute the anova table
   2. Check the significance of the difference in values between groups
   3. Plot the grouped boxplots

Relevant outputs for the anova table:

****

There are more than 95% that mean for quantitative var for those 3 categorical variables are the same.

My box plot aren’t grouped. I haven’t succeed at it. But they still looks good. You can find them in the notebook.

**Exercise 3**

Apply supervised learning on a quantitative outcome variable (recommended: Flow). With a linear regression model and a K nearest neighbors regressor (KNN) model

The result of my analysis between LinearRegression and KNN Regression is:



As expected, KNN is better (RMSE lower and r² adjusted higher).

APPENDIX

