How did we handle columns in match making?

**1. Age**

Before distribution

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Age is +vely skewed

Boxplot showing outliers above 58 so if age is above 58 then it becomes 58

After distribution

Chart, box and whisker chart

Description automatically generated

**2. Height - It was already normally distributed and with less outliers**

**Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated**

**3. Drinks**

Chart, histogram

Description automatically generated

If very often then rarely

If desperately then often

Chart

Description automatically generated

**4. drugs**

Chart, bar chart

Description automatically generated often to sometimes

Chart, bar chart

Description automatically generated

**5. Job**

Chart, bar chart, histogram

Description automatically generated

For merging refer to python file.(search for ”merging some similar jobs”)

Chart, histogram

Description automatically generated

**6. for pets column treatment refer to code – search for “handling pets column”**

**7. smoke**

**Chart

Description automatically generated**if when drinking then sometimes------ if trying to quit then sometimes

Chart

Description automatically generated

**8. Body Profile**

**Chart, histogram

Description automatically generated**

**Refer to code – search for “merging body profile to levels”**

**Chart, bar chart

Description automatically generated**

**9. Status**

**Chart

Description automatically generated**

**Single has way more values than other so except single merging all values to other**

**Chart, bar chart

Description automatically generated**

**10. interest**

**Chart, bar chart, histogram

Description automatically generated**

**Search for merging interest**

**Chart, bar chart

Description automatically generated**

**11. location**

#top 20 locations contributing to 87% of total count so other than top 20 will be "other"

top\_20\_locations=data['location'].value\_counts()[:20]