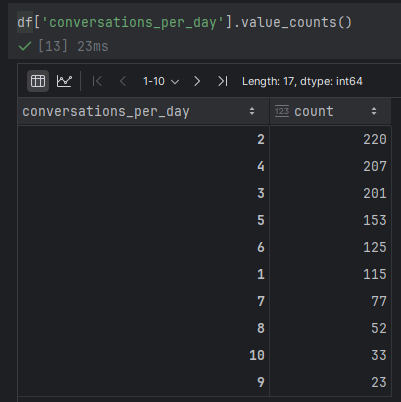
Part 1: Data Loading And First Look

1) 1250 rows and 25 columns

2)



I assume that it indicates the conversations per day the test subject has.

It’s ordinal because it’s categorical with a sort of natural order (less conversations correlates to higher count).

3)

|  |  |  |
| --- | --- | --- |
| Feature name | Description | Type |
| patient\_id | Patient identification number | Ordinal |
| age | Age of the patient | Ordinal |
| sex | Gender of the patient | Categorical |
| weight | Weight of the patient | Continuous |
| blood\_type | Patient’s blood type | Categorical |
| current\_location | Patient’s location | Categorical |
| num\_of\_siblings | Patient’s number of siblings | Ordinal |
| happiness\_score | Patient’s self-reported happiness score | Categorical |
| household\_income | Patient’s income on some scale. | Continuous |
| conversations\_per\_day | Num of conversations the patient has a day | Ordinal |
| sugar\_levels | Patient’s tested sugar levels | Continuous int |
| sport\_activity | How active the patient is | Categorical |
| pcr\_date | Date of the pcr test | Datetime |
| PCR\_01 | One result feature from the test | Continuous |
| PCR\_02 | Another result feature from the test | Continuous |
| PCR\_03 – PCR\_10 | Same here for all |  |

Partitioning the data:

4) It’s important to use the exact same split because we analyze the data to draw conclusions and we wouldn’t want them to be affected from changes in the underlying data but from our intentional processing steps.

Part 2: Missing Values and Outliers

5) train missing values:

Household\_income: 110

PCR\_02: 63

Test missing values:

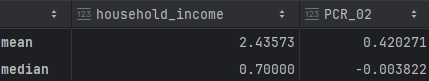
household\_income: 29

PCR\_02: 11

6) outliers are potential irregularities in the data. Usually, it’s defined as data that is not in the range of the lower bound and upper bound when they are:

This region should contain ~99.3% of the data if it’s normally distributed.

It seems there are outliers in the fields I plotted.

7)

There is a difference between the mean and median because the mean is more affected by the outliers and ‘pushed’ by them. I would prefer fillings the NANs with the median value as it closer to the majority of the data.

Part 3: Warming up with k-Nearest Neighbors

8)

A screenshot of a computer screen

Description automatically generated

It seems PCR\_01 and PCR\_08 make the data pretty seperetable.

9)



It doesnt contridict as each variable alone doesn’t need to corralate to the target spread (if it did we could come up with an easy classifier). The data can be sepertable in 2d and not in 1d.