ZS Datathon

BY-
Prakhar
IIT (BHU) Varanasi

Problems found in Data

- Presence of NA (missing values) in the Gender of Demographic data which needed to be imputed.
- Presence of many values corresponding to age >100 in age column.
- Presence of Undefined (U) value in the analysis data.

Data Preprocessing

- Used mode imputation for filling the missing values of Gender.

Justification: Presence of very few NA (<1%).

- Binned Age into 3 categories with a width of 40 each.

Justification: Take into account the beginners, professionals and retired doctors. Also confirmed from the distribution plot of Age that there is no trend between the intervals.

- Used groupby on Binned Age column for replacing ‘unknown’ in the value column.

Justification: Strong trend found between Binned Age and the value the variation by Chi squre test (p value=0), compared to other values like Region (p value of order 10^-5), Speciality_ID (p value of order 10^-120), Gender (p value of order 10^-83)

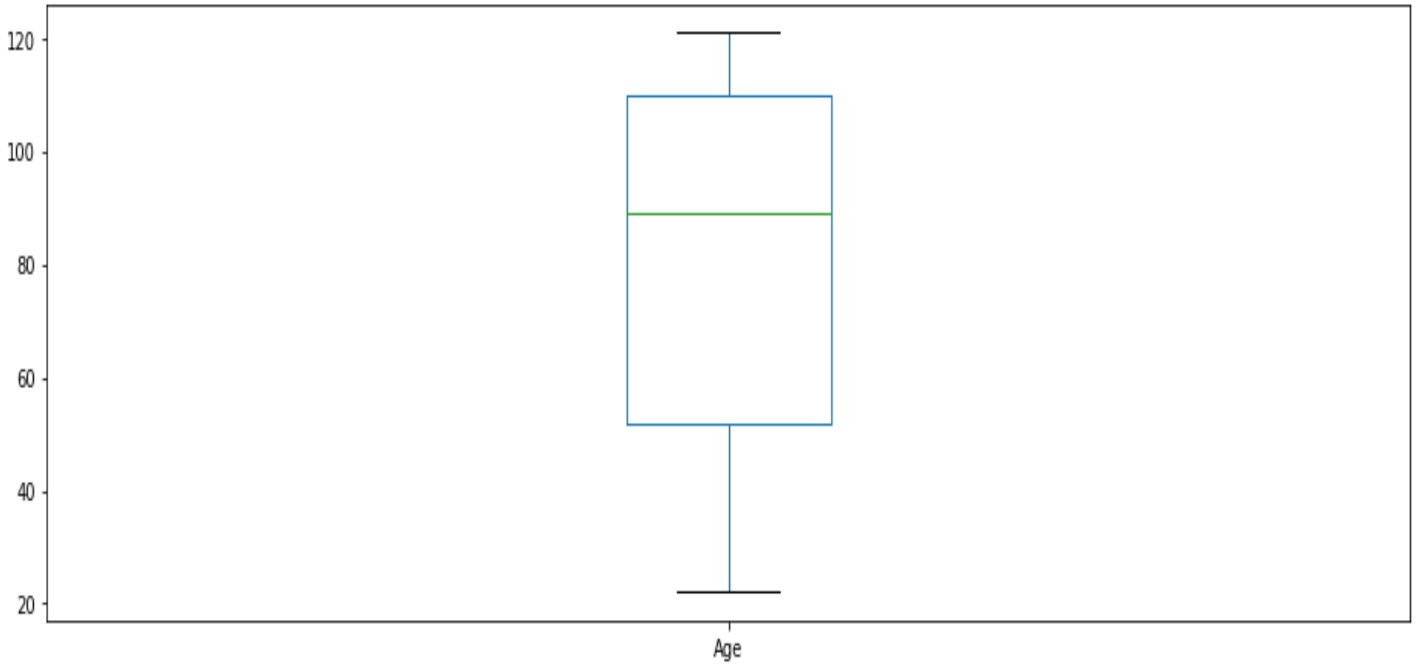

Fig:1 Boxplot of Age variable

Infered: Presence of no outlier in age

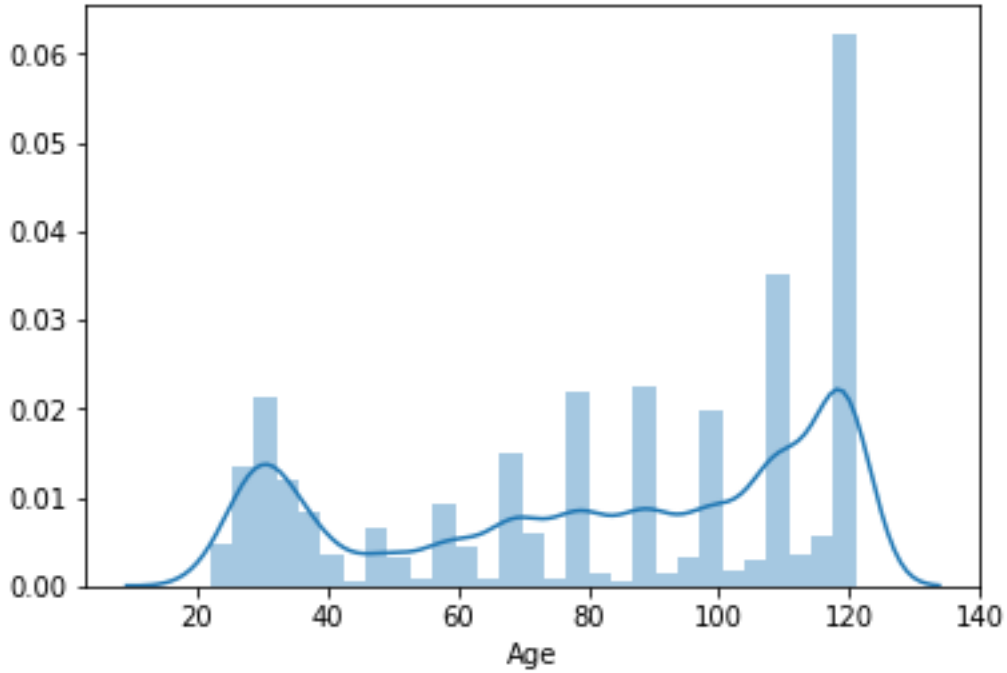

Fig:2 Distribution plot of age

Initial Approach

- Tried building a personalized model.
- A recommendation system by creating clusters of the rows using KNN.
- Steps followed :-
- Doing row-wise iteration, finding the highly correlated rows using the non misssing Affinity features(cont. variable).
- Applying KNN (k_neighbors=no. of rows/10) on the entire data (cont. + categorical variables) for highly correlated rows.
- Filling the missing value with the mean of the nearest neighbors.

Final Approach

- Majority of the predictors were categorical, so tree based method would suite the best to the data.
- High correlation between the pairs of Affinity features, so other features must also be taken into account while imputing values of a feature.
- No. of missing values in different features:-

Attribute	No of missing value
DMS	4379
DRT	4492
RL	4849
RR	9648
DEM	11040
OLA	11170
OLV	17027
P2P	29136

- Started missing value imputation with the feature containing minimum no. of missing values.
- Divided the data into train (not containing missing value) and test (containing missing value) set.
- Built the model on the training set using all the categorical variable and imputed continuous variable.
- Gave predictions using the model on the test set to fill the missing (NA) values.

THANK YOU