Raw data summery

2019 (data1_df_cp)
Total rows/ columns: 56,136/ 2,741
Columns used: 64
Rows after age removal: 42,739
Rows after error 85 removal: 41,950
Rows after removal of some responses:
Final dataframe rows/ columns: 24,584/ 64

2018 (data2_df_cp)
Total rows/ columns: 56,313/ 2,691
Columns used: 64
Rows after age removal: 43,026
Rows after error 85 removal: 42,252
Rows after removal of some responses:
Final dataframe rows/ columns: 24,575/ 64

2017 (data3_df_cp)
Total rows/ columns: 56,276/ 2,668
Columns used: 64
Rows after age removal: 42,554
Rows after error 85 removal: 41,761
Rows after removal of some responses:
Final dataframe rows/ columns: 23,961/ 64

2016 (data4_df_cp)
Total rows/ columns: 56,897/ 2,668
Columns used: 64
Rows after age removal: 42,625
Rows after error 85 removal: 41,844
Rows after removal of some responses:
Final dataframe rows/ columns: 23,613/ 64

2015 (data5_df_cp)
Total rows/ columns: 57,146/ 2,679
Columns used: 64
Rows after age removal: 43,561
Rows after error 85 removal: 42,752
Rows after removal of some responses:
Final dataframe rows/ columns: 23,810/ 64

Final total dataframe after merging (data_complete_df):
Rows/ columns: 120,543/ 64

Dataframe for the final prediction models:
Rows/ columns: 120,543/ 15