# Exploratory Data Analysis

The assignment is due on April 20 at 5 pm. The submission must include the source code, a readme file including how to run the code and specifying pseudo code for the relevant algorithms, and sample output. It must be done on github.

The data for this assignment is in data.csv with feature descriptions in feature\_descriptions.csv. These files are available in [share](https://nuwildcat-my.sharepoint.com/:f:/g/personal/dkl524_ads_northwestern_edu/EqSduFwR4C5Ltm9ZY5Fj9AABf4neSpHi8a7ElW1WiagWVA?e=gZ6H6t)

The final code must be embedded in a docker container (the repository must include Dockerfile) and the readme file must specify how to build in run the container (ideally there should be a run and build shell scripts). The code in the container must access the data from a local folder.

1. **Missing values and cleaning**

Analyze missing values. Remove columns with many missing values and for the remaining columns impute the missing values (argue the choice of the underlying imputation algorithm).

1. **Histograms and skewness**

Check skewness of each numerical feature and plot histograms for each such feature.

For each skew feature find a transformation that makes it (almost) normal.

Create a dataframe with transformed features.

1. **Box plots**  
   Consider the same numerical features with the appropriate transforms (if any) from the previous question.

Generate boxplots of these features, and identify and remove the outliers.

Analyze the categorical feature TARGET. Repeat the above part of this question but generate separate boxplots for "TARGET"=0 and "TARGET"=1. Which features have significant differences among the two groups?

Take a closer look at the "AMT\_INCOME\_TOTAL" feature (with the appropriate transform. Generate separate boxplots for different levels of education ("NAME\_EDUCATION\_TYPE"). Arrange the education levels in the following order:  
"Lower secondary" < "Secondary / secondary special" < "Incomplete higher" < "Higher education".   
What do you observe? Are there any outliers within each group.

1. **Bar plots**

Generate a bar plot to count the number of applicants with different housing types ("NAME\_HOUSING\_TYPE). Repeat this but now separate the counts by family status (NAME\_FAMILY\_STATUS) within each group.  
Draw observations.

1. **Feature engineering**
   1. Create a new column AGE from DAYS\_BIRTH by dividing the entries by 365.
   2. Create a new column AGE\_GROUP depending on the AGE:  
      - "Very\_Young": 19-25  
      - "Young": 25-35  
      - "Middle\_Age": 35-60  
      - "Senior\_Citizen": >60
   3. Plot the proportion of applicants with "TARGET"=1 within each age group. What do you observe?
   4. Repeat (c) but now separate the proportions by gender (CODE\_GENDER) within each age group.