* Author: Petra Besenhard
* Project Title: "Exploring skill distribution tails through stochastic dominance"

DATA GENERATION:

All the data for this replication paper were generated using Eeckhout, Pinheiro, and Schmidheiny's (2014)*** Stata code for their 'Spatial Sorting' Paper.
The Stata code is available through the Jounral of Political Economy data archive (https://www.journals.uchicago.edu/doi/suppl/10.1086/676141). The code was followed exactly to generate the data for this paper, in order to guarantee an exact replication of the results.

ANALYSIS:
The analysis for this paper was complete in R.
The following R packages were used: quantreg, ggplot2, EnvStats, car

Datasets:

skillacs.csv: Allows analysis of inidvidual skills from 2009 ACS (*originial data comes from IPUMS USA and requires a user registration)
skillcps.csv: Allows analysis of individual skills from 2009 CPS

(The final dataset skillcps.csv is available in conjunction with this readme file, while the skillacs.csv dataset can be obtained by
downloading the raw data from IPUMS USA and following Eeckhout et al.'s Stata code for the creation of the final dataset.)

Dataset details:

##skillacs.csv##
*Creating this dataset requires registration with IPUMS USA in order to download the raw ACS data.

Total number of observations: 655,961
Total number of variables: 61
Variables of importance for replication and stochastic dominance analysis (*Eeckhout et al.'s different utility measures were utilized in the stochastic
dominance analysis):
- pop2009: Population size for each CBSA area
- cbsa: Indicates CBSA area (Core-based statistical area)

- rentindexcbsa: Index created to observe housing costs as a representation of cost of living
- wage: Individual's wage
- lwage: Individual's log wage
- lutility1:Cobb-Douglas utility based on CBSA rent index from ACS hedonic housing prices, baseline
- lutility3: Alternative Cobb-Douglas based on price index of PUMA regions

**NOTE: In order to generate this dataset, the following additional raw data needs to be downloaded to utilize Eeckhout et al.'s Stata code:

- cbsa: CBSA labels from the Missouri Census Data Center (MCDC)
- pumacbsa: PUMA to CBSA correspondance from MCDC
- nectacbsa: NECTA to CBSA correspondance from MCDC
- metrodivcbsa: Metro Division to CBSA correspondance from MCDC
- cbsa-est2009-01: CBSA population from Census

##skillcps.csv##
Raw data is downloaded from the 2009 CPS (morg09 from NBER's Merged Outgoing Rotation Groups). Details on how this final dataset was created can be
found in Eeckhout et al.'s Stata code.

Total number of observations: 103,696
Total number of variables: 51
Variables of importance for replication and stochastic dominance analysis (*Eeckhout et al.'s different utility measures were utilized in the stochastic
dominance analysis):

- pop2009: Population size for each CBSA area
- cbsa: Indicates CBSA area (Core-based statistical area)
- rentindexcbsa: Index created to observe housing costs as a representation of cost of living
- wage: Individual's wage
- lwage: Individual's log wage
- grade92: Individual's schooling level
- occ00: Occupation categories
- ind02: industry categories
- age: Individual's age
- foreign: Indicates whether individual is native or foreign-born
- lutility1: Cobb-Douglas utility based on CBSA rent index from ACS hedonic housing prices, baseline
- lutility2: Alternative Cobb-Douglas utility based on ACS hedonic housing prices, price of top 10% PUMA
- lutility4: Stone Geary utility based on ACS hedonic housing prices

>>>>> The R scripts for the replication section as well as the stochastic dominance analysis are available upon request.

***Eeckhout, J., R. Pinheiro, and K. Schmidheiny, 2014: Spatial Sorting. *Journal of Political Economy*, 122, 3, 554-620.