Look into the negative values

* If the household have e.g (-9, 20.2) -> 20.2
* Appropriate skip -> 0
* Mean/Median for thos who don’t know
* Look at those with big proportion (20%)
* **> 15% drop**
* **Check the variables and do aggregation and categorisation**
* **- vehcount**

**MSASIZE**

**Jiaqi**

**BIKESHARE** Count of Bike Share Program Usage – average (drop due to more than half appropriate skip)

**CARRODE** Count of People in Vehicle to Work – average (drop)

**CARSHARE** Count of Car Share Program Usage – average (recode to 0 if less than 0)

**CNTTDTR** Count of person trips on travel day - average

**DISTTOSC17** Road network distance, in miles, between respondent's home address and school address – average, have to check if got student in the family (drop due to majority are not students)

**FLEXTIME** Flex Time – ratio (1-yes, 2 – no) ~~–~~ **~~remove~~** (recode those <0 to new categorical value 0 means unknown, will do one hot encoding)

**GT1JBLWK** More than One Job – COUNT/max? (recode those <0 to new value 0)

**HEALTH** Opinion of Health – average (higher is better) (recode those <0 to new categorical value 0 means unknown and then choose the highest number)

**HHSIZE** Count of household members

**MCUSED** Count of Motorcycle or Moped Trips - AVERAGE (drop due to majority data are appropriate skip)

**NBIKETRP** Count of Bike Trips – AVG (recode those <0 to 0 and then apply average)

**WKFMHMXX** Count of Days Worked from Home in Last Month (drop due to more than 90% chose appropriate skip, there is no obvious link between appropriate skip and worker/student status)

**Wei Ji**

**NOCONG** Trip Time in Minutes to Work without Traffic - AVG

**TIMETOWK** Trip Time to Work in Minutes - AVG

**NUMADLT** Count of adult household members at least 18 years old

**~~NWALKTRP~~** ~~Count of Walk Trips – AVG~~ **no need as already mentioned in household set**

**OCCAT** Job Category – COUNT

**~~PTUSED~~** ~~Count of Public Transit Usage - AVG~~ **no need as already mentioned in household set**

**PUBTIME** Minutes Spent Transferring to Work – AVG

**~~RIDESHARE~~** ~~Count of Rideshare App Usage – AVG~~ **not useful**

**~~R\_AGE\_IMP~~** ~~Age (imputed) – AVG - done~~

**~~R\_SEX\_IMP~~** ~~Gender (imputed – male%, female%(adults) – done~~

**WKRMHM** Option of Working from Home

**WRKTIME** Arrival Time at Work

**WKFTPT** -> ratio

**DISTTOWK17** Road network distance, in miles, between respondent's home address and work address – average -> have to check if got worker

**~~EDUC~~** ~~Educational Attainment – average (higher is better) - done~~

Later – one hot encoding/bin