```
#after wget, put data in HDFS

hdfs dfs -mkdir relate

hdfs dfs -put surveyrelate_nohead.csv relate

#put data into HIVE

create external table relate
(
id string,
name string,
survey_score int,
rating int
)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
    "separatorChar" = ",",
    "quoteChar"     = '"',
    "escapeChar"    = "\\"
    )
STORED AS TEXTFILE
LOCATION '/user/w205/relate'
;

#clear out invalid characters in data

create table relate_clean AS SELECT
id, name, survey_score, rating FROM relate
WHERE survey_score <> "Not Available" AND
rating <> "N"
;

#run correlation

select corr(survey_score, rating) FROM relate_clean;
```