**Homework #1**
**14-March**

Exercise 1 (25pts)
Using the 'u.data' file (included), answer the following questions. You'll need to load the
data in a dataframe using the following code fragment. Write your code to answer the question.
**An answer without code will receive 0 points.**

```
# dataset schema user id | item id | rating | timestamp
schema = "userid INT, itemid INT, rating INT, timestamp LONG"
df = spark.read.format("csv") \
        .option("delimiter", "\t") \
        .schema(schema) \
        .load("u.data")
# Convert timestamp to a date
df.select(to_timestamp("timestamp")).show()
```

(a) How many records are in the dataset?

(b) How many 5-star ratings are in the dataset?

(c) What is the average rating?

(d) Which user has the highest number of ratings?

(e) What are the minimum and maximum ratings?

(f) What is the average rating for the user with userid equal to 429.

Exercise 2 (25pts)
Using the 'departuredelays.csv' file (included), answer the following questions. You'll need to load the data in a dataframe first. Write your code to answer the question. **An answer without code will receive 0 points.**

(a) How many records are in the dataset?

(b) Which flight had the longest delay?

(c) Show all flights between San Francisco (SFO) and Chicago (ORD) with at least an 1 ½ hour delay

(d) How many flights had delays between 2 and 5 hours?

(e) What is the average delay? Distance?

Exercise 3 (50pts)
Using the 'us_population.json' file (included), answer the following questions. You'll
need to load the data in a dataframe first. Write the code that you used to answer the question.
**An answer without code will receive 0 points.**

(a) What is the population of Massachusetts (MA)?

(b) What is the <u>city</u> with lowest population? Highest population?

(c) Which cities have a population that is greater than the average population?

(d) Which <u>state</u> has the highest population? Lowest population?

(e) What is the average population in each state?