**Here is the main steps you have to follow in order to clean the data efficiently**

**General notes:**

1. Detect any duplicates and save them into a separated CSV file
2. Do not try to deal with missing values in this stage. We will leave them as is for now, and will fill them with the most appropriate value after cleaning all the features in the dataframe.
3. Correct the types of all features, for example we expect the area to be ‘int’ not ‘object’

**Start cleaning the features in the following order:**

1. **location**:
   1. Extract geographic information such as governorate and city/district
   2. If the city/district is missing, fill it with the name of the governorate
   3. Create a new feature and name it ‘region’. It is meant ot group the governorates into 6 categories:
      1. **Capital:** Cairo and Giza
      2. **Alexandria:** Alexandria
      3. **Upper Egypt:** Fayoum, Menia, Bani-Suef, Assyout, Sohag, Qena, Luxor, Aswan
      4. **Lower Egypt:** Qalyoubia, Dakahlia, Gharbia, Behera, Menoufia, Kafr-El Sheikh, Dumiat, Sharqia
      5. **Canal:** Ismailia, Port-Said, Suez
      6. **Border:** Bahr-Ahmar, South Sinai, North Sinai, New Valley, Matrough
2. **title:** leave it as is for now
3. **bedrooms, bathrooms, level:** we have two options
   1. Leave them as is (categorical features) to take into account **10+** values.
   2. Convert them into integer: in this case you will need to delete the plus sign from 10+ and treat it as if it is 10
4. **area:**
   1. Convert it to numeric: integer.
   2. The feature has unreasonable values. Therefore, we will work on outlier detection later on.
5. **Furnished:** The feature has missing values, try to fill them with the correct values based on the description or the title.
6. **type:** it seems to be nice and clean, if so leave it as is
7. **amenties:** 
   1. Create a new feature and name it “**amenties\_specified**. It should take 0 if “**amenties”** is missing and 1 otherwise.
   2. Split the **amenties** into dummy features (This step would have been much easier if addressed during data collection)
   3. Delete the original feature
8. **compound:** 
   1. Create a new feature and name it “**in\_compound**. It should take 1 if “**وند”** is in title or description
   2. If the value of ‘**in\_compound”** is 1 and the value of “**compound”** is missing replace the value of “**compound”** with **“compound name not specified”**
   3. If the “**compound”** feature still has missing values, fill them with “Not Available” or “Not in compound” or any other value based on your best judgment.
9. **date:**
   1. Extract: Year, Month, Day of the week.
   2. Store each of them into a separate feature
10. **delivery date:**
    1. Create a new feature and name it “**delivery\_date\_specified**. It should take 0 if “**delivery\_date”** is missing and 1 otherwise.
    2. Fill in the missing values using “Not Specefied”

**Outliers:**

1. Plot the numerical features against each other (price against area) and try to spot any inconsistencies.
2. Do not forget to take into account all other features while you are performing step 1.