Practice Exam III

Questions

The dataset in `Real estate valuation data set.xlsx` contains real estate prices in Sindian (Xindian) District of New Taipei City in Taiwan. The seven inputs are as follows:

Column name	Description
X1	transaction date. (for example, 2013.250=2013 March, etc.)
X2	age of the house. (unit: years)
X3	distance to the nearest MRT station. (unit: meters)
X4	number of convenience stores within walking distance.
X5	latitude. (unit: degree)
X6	longitude. (unit: degree)

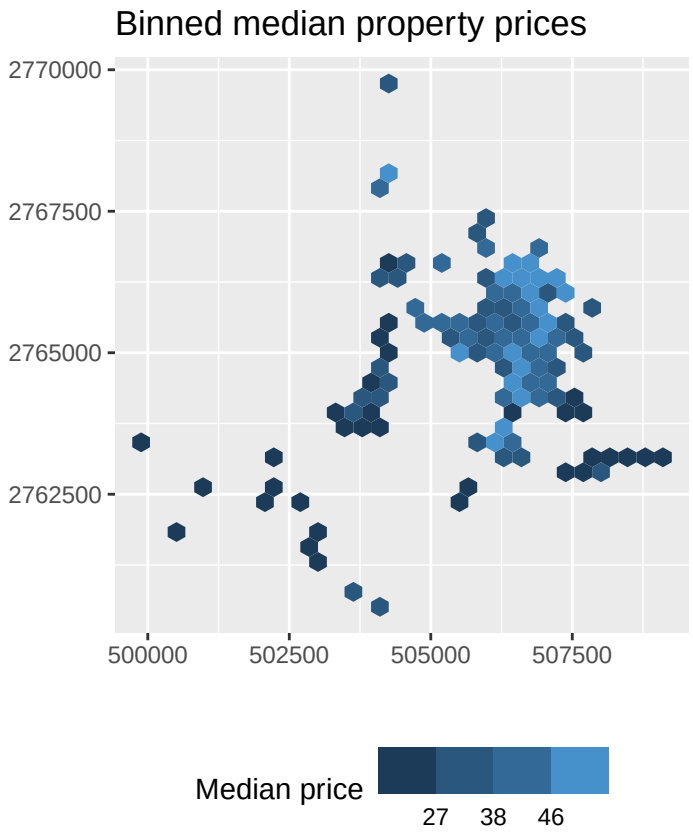
The output variable, price, is as follows:

- Y = house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Question 1 (15 marks)

Re-create the following plot as closely as you can. You may need the following information:

1. The EPSG code 3825 was used to project the coordinates.
2. If the hex bins are not available to you, use `stat_summary_2d`.
3. The fill aesthetic was mapped the the median house price in each bin.



Question 2 (15 marks)

A property analyst informs you that the centre of town is at (506800, 2765500). Add a column to your dataset that contains the Euclidean distance the the centre of town.

Now, create **two** graphics that demonstrate how price varies with the other variables. Please ensure that the following criteria are satisfied: * Distance to the centre of town must be present in at least one of the graphs. * The price variable must be present in both of them.

Question 3 (10 marks)

The same property analyst suspects that price has a non-linear relationship with house age: For modern houses, the price decreases with age but for vintage houses, the prices increases with age.

The problem is that the change-point is unknown - at what age should we classify a house as vintage? The following function will return the parameters for a line of best fit through these points, given a change-point.

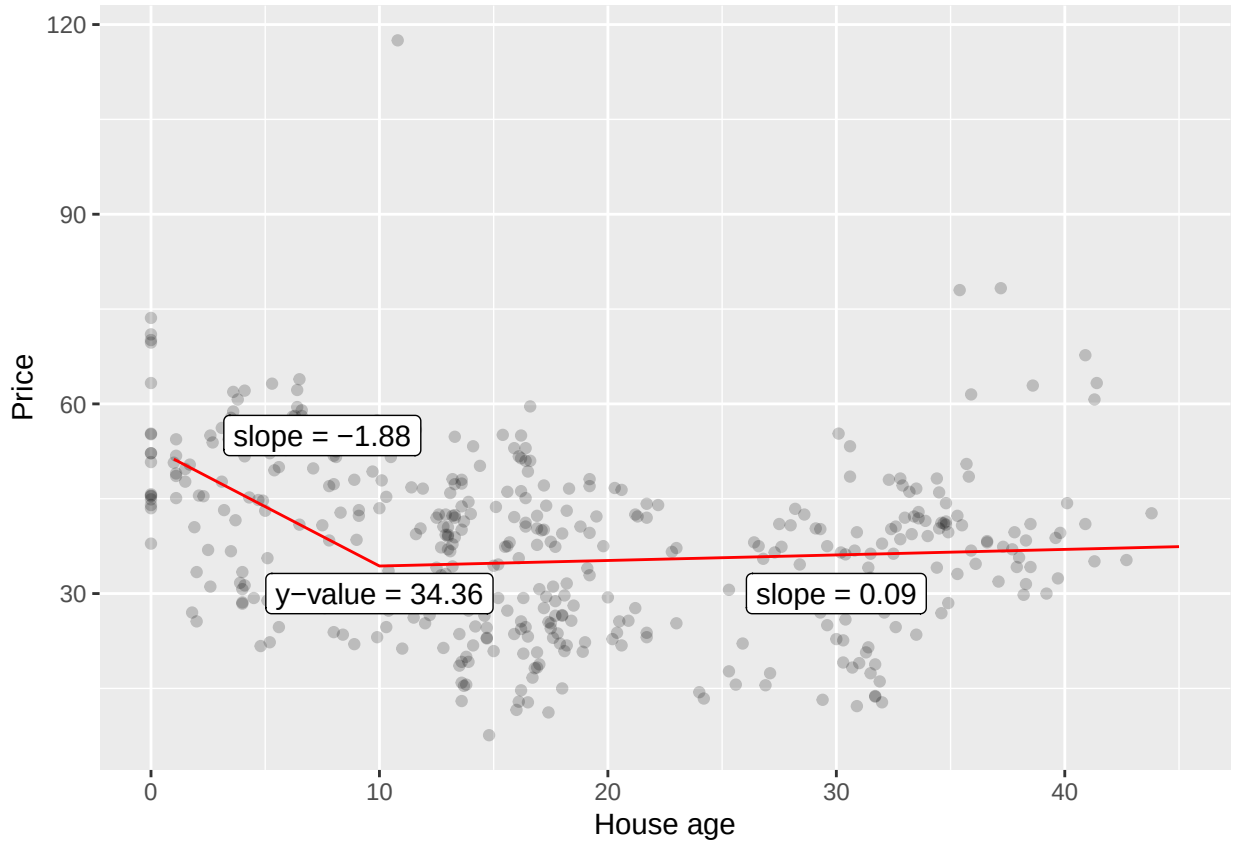
```
get_gradient_intercept <- function(price, age, breakpoint) {
  modern <- pmin(age - breakpoint, 0)
  vintage <- pmax(age - breakpoint, 0)
  lm0 <- lm(price ~ modern + vintage)
  coef(lm0)
}
```

For instance, *if the analyst believed* the change-point was at 10 years:

```
(coefs <- get_gradient_intercept(re2$price, re2$house_age, 10))
```

```
## (Intercept)      modern     vintage
## 34.35780172 -1.87805087  0.08756953
```

Then the fitted line would be plotted like this:



Find the **optimal change-point**. Desribe what criteria you used to find it, and then create a properly annotated plot similar to the one above that summarises the relationship.