Ecole Nationale Supérieure de Statistiques et d'Economie Appliquées (ENSSEA)

Examen R

Dr. ASRI

2023-01-26

Contents

Problème

Les prix des maisons est un problème réccurent en économie. Notre problème est d'analyser les prix des maisons en Californie (USA) en 1990. La base de données contient des informations aggrégés sur les maisons en rues (city block) califoriennes. Cela implique que chaque ligne représente les informations dans toute la rue citée.

Les variables de cette base de données sont :

longitude la longitude

latitude la latitude

la longitude et la latitude sont des mesures pour l'emplacement dans la carte (longitude : est/ouest et latitude nord/sud)

housing_median_age l'age médiane des maisons dans cette rue

total_rooms nombre total de chambres (cusuines et salle de bains inclus) des maisons dans cette rue

total_bedrooms nombre total de chambres (chambres seulement) des maisons dans cette rue

population nombre total de la population de cette rue

households nombre totales des ménages dans cette rue

medianIncome le revenu médian des ménages dans cette rue (mesuré en 10000$)

medianHouseValue valeur médiane de la maison dans cette rue (mesuré en $)

oceanProximity proximité à l'océan.

Travail à faire

Charger le package tidyverse et lisez la base données "data.csv" dans un objet qui doit être nommé **data** et répondez aux questions suivantes :

Question 01 .

1. Préciser le nombre de lignes de cette base de données.

2. donner le sommaire statistiques des variables numériques

3. préciser les différentes catégories des variables catégorielles.

4. préciser les variables qui contiennent des valeurs manquantes et le nombre pour chacune de ses variables.

5. est ce que le nombre de valeurs manquantes est considérable ? justifier.

Question 02.

Exécuter le code suivant en changeant la valeur du "matricule" par votre propre matricule (le fichier excel contenant les matricules est envoyé avec ce document).

```
set.seed(as.numeric("matricule"))
data <- data %>%
  slice_sample(n = 5000)
```

par exemple si un étudiant a un code : 2031063500

```
set.seed(as.numeric("2031063500"))
data <- data %>%
  slice_sample(n = 5000)
```

1. Expliquer que fait ce code exactement ?

2. Quel est le rôle principale de cette opération ?

PS. Utiliser cette nouvelle base pour le reste des questions

Question 03.

1. recalculer les sommaires statistiques pour les variables numériques.

2. y-a-t'il un changement par rapport à la première question ? pourquoi ? expliquer.

3. tracer l'histogramme de chaque variable numérique

4. essayer de proposer une méthode pour dessiner les histogrammes de toutes les variables en un seul graphique (ne pas utiliser patchwork)

5. comparer le graphique de la question précédente avec un code qui utilise patchwork. (comparer le résultat et le code)

6. Analayser les variables individuellement et essayer de détecter d'éventuelles anomalies.

Question 04.

1. Calculer l'age médian des maisons dans toute la californie.

2. calculer la médiane de la valeur des maisons californiennes proche de l'océan pour les ménage avec revenu moins de 45000$

3. calculer la variance et l'écart type du nombre de chambres des rues californiennes pour les rues avec population supérieur à 10000 personnes et qui sont moins 1h de l'océan.

4. donner la fréquence (le nombre de rues) et la fréquence relative des rues ayant un revenu médian entre 25000 et 45000 dollars et qui ait le nombre totale de chambres (tous inclus) supérieur à 500

5. Calculer l'erreur absolue moyenne de nombre de chambres (chambres seulement) pour chacune des catégories des maisons par rapport à leurs proximité à l'océan.

6. Calculer les déciles des ménages pour les rues proche de la baie (bay) et avec population supérieur à 1000

7. Calculer les 30%,60% et 85% quantiles des valeurs des maisons pour les rues avec un nombre de ménage inférieur à 100

8. calculer les quartiles des valeurs des maisons pour chaque proximité de l'océan. Commenter.

Question 05.

1. Tracer le nuage de points de latitude vs longitude ? que remarquez vous ?

2. Comparer la forme générale de ce nuage de points et la carte de la californie. Expliquer.

3. y-a-t'il un coté qui a approximativement une fréquence plus ou moins élevé ?

4. tracer le même nuage de points de la question 5.1 et colorer par la proximité de l'océan.

5. tracer le même nuage de points de la question 5.1 et colorer par la valeur des maisons. expliquer ce graphique

6. y-a-t'il une différence entre les deux graphiques 5.4 et 5.5 ? proposer une méthode pour combiner les deux informations en un seul graphique.

7. pour répondre d'une manière exacte à la question 5.3, on propose de créer une nouvelle variable."coté". En utilisant longitude et latitude on va créer 4 catégories.

Proposer 4 catégories pour séparer la figure en 4 cotés différents (il n'existe pas de solution exacte, cela va dépendre de la vision de l'étudiant).

8. tracer le nuage de point de points de la question 1 et ajouter les séparations porposer en question précédentes 5.7 (utiliser **geom_hline** et **geom_vline** avec 4 couleurs)

9. combiner les idées de graphiques 5.7 et 5.8 pour tracer un graphique contenant toutes ses informations.

10. Commenter le graphique précédent (5.9 ou 5.7 si vous n'arrivez pas à tracer le graphique 5.9) et essayer de tirer des conclusions.

Question 06.

1. tracer un graphique de de la distribution de la proximité à l'océan (fréquence)

2. tracer la boite à moustaches des nombres de chambres (les deux variables)

3. tracer la boite à moustaches des nombres de chambres en distinguant les différentes proximités à l'océan.

4. tracer l'histogramme et la densité du revenu médian en distinguant entre les différents cotés de la californie ? (si la variable n'est pas été crée utiliser la proximité à l'océan)

5. tracer le nuage de points de la valeur médiane de la maison vs l'age des maisons

6. tracer le nuage de point précédent en distinguant entre la proximité à l'océan. Commenter.

7. tracer le nuage de points de la valeur médiane de la maison vs nombre de chambres (les deux variables). Puis colorer par la proximité de l'océan et puis colorer par le coté (si c'est possible).

8. tracer le nuage de points de population vs ménage. Commenter .

Question 07.

1. Calculer la corrélation entre la valeur médiane de la maison et les autres variables numériques.

2. tracer le graphique de la corrélation entre toutes les variables numériques

3. calculer le sommaire statistique de la médiane de la valeur de la maison par rapport à chaque coté et par rapport à la proximité à l'océan.

Question 08.

Etablir une conlusion générale sur les différentes variables dans la base de données en précisant les différentes anomalies et les différentes relations entre les variables et la valeur médiane de la maison.

Essayer de donner un titre à chaque catégorie de questions.

Instructions

1. L'étudiant doit envoyer un ou plusieurs fichiers .Rmd (Rmarkdown) ou .R un r script contenant les différents codes et réponses. Le fichier doit obligatoirement contenir le nom, prénom et le groupe de l'étudiant.

2. Chaque question doit être répondu soigneusement et chaque graphique doit être interprété.

3. L'étudiant est encourage a enrichir l'analyse et l'esthétique générale des graphique et cela va apporter des points supplémentaires.

4. Les réponses similaires vont être considérées comme des tentatives de triches et vont être sanctionnées.