PROJET 1 :

Analyse des données du marché du logement dans le projet R

Effectuer des analyses EDA, des visualisations de données, des tests d'hypothèses et des régressions linéaires dans R sur les données du marché immobilier

Partie 1 : Analyse des données

1. **Installer et charger les bibliothèques** . Avant d'effectuer une analyse, il est essentiel de s'assurer que les bibliothèques nécessaires sont installées et chargées. Dans ce code, vous devez exécuter la première ligne si vous n'avez pas installé les packages. Sinon, passez directement à la deuxième ligne :

    a. **install.packages(c("ggplot2", "dplyr"))** – Cette ligne de code installe deux packages R essentiels : ggplot2 pour la visualisation des données et dplyr pour la manipulation des données.

    b. library(ggplot2) et – Ces commandes chargent les bibliothèques et , rendant leurs fonctions disponibles dans les étapes suivantes. **library(dplyr)** ggplot2 dplyr

2. **Chargement des données et lecture du fichier CSV**

data <- read.csv('path_to_file') – Cette commande lit un fichier CSV contenant l'ensemble de données sur le logement et le stocke dans une variable nommée data .

2. **Exploration initiale des données**

- **Statistiques récapitulatives**

print(summary(data)) – Cette commande calcule et imprime un résumé de chaque variable de l'ensemble de données, y compris généralement des mesures telles que le minimum, le premier quartile, la médiane, la moyenne, le troisième quartile et le maximum pour les variables numériques. Il s'agit d'un premier aperçu de la distribution des données.

- **Affichage des premières lignes**

head(data) – Cela fournit un instantané des six premières lignes de l'ensemble de données, donnant un aperçu de sa structure et du type de données qu'il contient.

- **Affichage des dernières lignes**

tail(data) – Semblable à head() , cette commande affiche les six dernières lignes de l'ensemble de données.

3. **Calcul de statistiques descriptives**

- **Fonction de mode personnalisé**

mode_func() – Étant donné que R n'a pas de fonction intégrée pour le mode, une fonction personnalisée nommée mode_func() est définie pour calculer le mode d'un vecteur numérique.

- **Calcul des statistiques descriptives**
    - means , modes , medians , et sds – Ces commandes appliquent leurs calculs statistiques respectifs à chaque colonne ou variable de données.
    - print(data.frame(...)) – Cette ligne crée un nouveau bloc de données qui consolide les statistiques calculées (moyenne, mode, médiane et écart type) et l'imprime pour révision.

4. **Analyse de corrélation**

- **Calcul de la matrice de corrélation**

correlation_matrix <- cor(data, use = "complete.obs") – Cette commande calcule les corrélations par paires entre chaque variable de l'ensemble de données et stocke les résultats dans une matrice nommée **correlation_matrix** .

- **Trouver la corrélation maximale**

max(correlation_matrix[correlation_matrix < 1]) – Cette ligne identifie le coefficient de corrélation le plus élevé dans la matrice, à l'exclusion de toute valeur de 1 (corrélation parfaite). Cela peut être utile pour identifier la paire de variables les plus étroitement liées, en dehors des autocorrélations.

5. **Imputation des données manquantes**

- **Remplissage des valeurs manquantes**

data <- data %>% mutate_all(...) – En utilisant le dplyr mécanisme de canalisation du package (**%>%**), cette commande met à jour chaque colonne du data jeu de données. Si une valeur est manquante (NA), elle la remplace par la moyenne de cette colonne. L'opération garantit que les données manquantes ne gênent pas les analyses ultérieures.

Conclusion : Ce code est un script complet d'exploration initiale et de prétraitement des données du marché immobilier. Il garantit que les données sont comprises, nettoyées et prêtes pour des tâches d'analyse ou de modélisation plus approfondies.

Partie 2 : Visualisation des données

Nous allons créer un histogramme présentant la distribution à Median.Home.Value l'aide du package ggplot2 dans R.

Description du code

1. **Affectation du tracé à une variable**

histogram <- ...

L'intrigue entière, construite avec toutes les couches et modifications mentionnées ci-dessus, est stockée dans la histogram variable.

2. **Initialiser l'intrigue avec des données et de l'esthétique**

ggplot(data, aes(x = Median.Home.Value))

- ggplot():Cette fonction lance la création d'un tracé à l'aide du package ggplot2.
- data:Il s'agit du bloc de données contenant les données que vous souhaitez visualiser.
- **aes(x =** Median.Home.Value**)**:Dans la aes() fonction (abréviation de « esthétique »), vous spécifiez que l'axe des x sera basé sur la Median.Home.Value variable du bloc **de données** .

4. **Ajouter une couche d'histogramme**

geom_histogram(bins = 9, fill=darkcyan", color="white", alpha=0.7)

- geom_histogram() :Cette fonction ajoute une couche d'histogramme au tracé.
- bins = 9 : Cela spécifie que les données doivent être divisées en 20 compartiments (intervalles). Cela peut affecter la granularité de la visualisation.
- fill="darkcyan" :Cela définit la couleur de remplissage des barres de l'histogramme sur une nuance cyan foncé.
- color="white" :Cela définit la couleur des bordures autour de chaque barre de l'histogramme comme blanche.
- alpha=0.7 : Cela définit la transparence des barres. Une valeur de 1 signifie totalement opaque, tandis que 0 signifie totalement transparent. Ici, 0,7 offre un équilibre, permettant des visualisations potentiellement superposées.

5. **Étiquetage de l'intrigue**

labs(title = "Distribution of Median Home Values", x = "Median Home Value", y = "Frequency")

- **labs() :Cette fonction permet de personnaliser les étiquettes du tracé.**
- **title :Ceci définit le titre principal de l'intrigue.**

- **x :Il s'agit de l'étiquette de l'axe des x.**
- **y :Il s'agit de l'étiquette de l'axe des y.**

6. **Appliquer un thème**

theme_minimal()

Cette fonction applique un thème minimaliste à l'intrigue, ce qui supprime certains éléments d'arrière-plan et lignes de grille, ce qui donne un aspect plus net.

L'histogramme produit offre un aperçu de la distribution des valeurs médianes des maisons de l'ensemble de données. En visualisant les données de cette manière, on peut rapidement discerner des tendances, des anomalies ou des caractéristiques essentielles des valeurs des maisons dans l'ensemble de données étudié. Les choix de couleurs et le thème spécifiques améliorent la clarté et l'attrait visuel du graphique, mais sont facultatifs, alors n'hésitez pas à ajouter votre propre formatage personnalisé.

Partie 3 : Tests d'hypothèses

Décomposons la solution du test d'hypothèse :

1. **Type de test :** Nous avons effectué un test t de Student, qui compare les moyennes de deux échantillons indépendants, en supposant des variances égales.
2. **Données :** Nous avons testé la variable Median.Home.Value sur deux catégories, élevée et faible, au sein de la Crime.Category variable.
3. **Statistique de test** : **t = 1,0681**
   La statistique t mesure la différence entre les moyennes des groupes concernant les erreurs standard. Une valeur t proche de 0 suggère que les deux groupes ont des moyennes similaires, tandis qu'une valeur t absolue plus élevée indique une différence entre les moyennes des groupes.
4. **Degrés de liberté** : **df = 479**
   Les degrés de liberté (df) indiquent le nombre de valeurs indépendantes qui peuvent varier dans le calcul. Pour un test t à deux échantillons avec des variances égales, les degrés de liberté sont calculés comme suit :

df=n1+n2−2,Danshetlet n1 unnd n2 unlet thet munmplet mjeAvecetm lef thet tDansle glleda nspmdf=n1+n2−2,where n1 and n2 are the sample sizes of the two groups

5. **Valeur p** : **valeur p = 0,286**
   La valeur p nous aide à décider s'il faut rejeter l'hypothèse nulle. Une petite valeur p (généralement ≤ 0,05) indique que vous pouvez rejeter l'hypothèse nulle. Dans ce cas, la valeur p est de 0,286, supérieure à 0,05. Cela signifie qu'il n'y a pas

suffisamment de preuves pour rejeter l'hypothèse nulle selon laquelle les moyennes des deux groupes (catégories de criminalité élevée et faible) sont égales.

6. **Hypothèse alternative :** La **véritable différence de moyennes entre les groupes à criminalité élevée et faible n'est pas égale à 0.**
   Cela indique une véritable différence de moyennes entre les catégories de criminalité élevée et faible.

7. **Intervalle de confiance à 95 %** : **-0,5635464 1,9058320**
   Cet intervalle estime où se situe la véritable différence de moyenne entre les deux groupes avec un niveau de confiance de 95 %. Étant donné que cet intervalle contient 0, il est cohérent avec la possibilité que les groupes n'aient aucune différence de moyenne.

8. **Exemples d'estimations**

- **Moyenne du groupe High : 47,69264**
- **Moyenne du groupe Low: 47,02150**
- Il s'agit des moyennes d'échantillons pour les Median.Home.Value catégories de criminalité élevée et faible, respectivement. La différence entre ces moyennes est 47,69264−47,02150=0,6711447,69264−47,02150=0,67114.

Interprétation

Le test t a été effectué pour comparer les valeurs médianes des logements entre les catégories de criminalité élevée et faible. La valeur p obtenue de 0,286 suggère qu'il n'y a pas de différence statistiquement significative dans les valeurs médianes des logements entre les deux catégories de criminalité au niveau de signification conventionnel de 5 %. Les moyennes de l'échantillon montrent une différence de 0,67114 unité, ce qui n'est pas statistiquement significatif. L'intervalle de confiance à 95 % confirme encore cela en couvrant les valeurs négatives et positives (−0,4490 à 1,5093), ce qui indique que nous ne pouvons pas être sûrs de la direction de la différence sur la base de cet échantillon.

D'après les données, il n'existe pas suffisamment de preuves statistiques pour suggérer une différence significative dans la valeur médiane des maisons entre les catégories de criminalité élevée et faible.

En termes plus simples, les données ne fournissent pas de preuves solides suggérant que la catégorie de criminalité (élevée ou faible) a un impact sur la valeur médiane des maisons, du moins pas dans une mesure statistiquement significative.

Partie 4 : Régression linéaire

Cette étape analytique vise à établir des relations prédictives entre les variables. À l'aide de techniques telles que la régression linéaire, nous cherchons à quantifier la manière dont

certains facteurs, comme le nombre moyen de pièces d'une maison, influencent la variable de résultat, dans ce cas, la valeur médiane d'une maison. En comprenant ces relations, nous pouvons faire des prévisions éclairées, tirer des enseignements et fournir des recommandations concrètes en fonction de la force et de la nature de ces relations.
Vous trouverez ci-dessous un guide complet pour effectuer une analyse basée sur R.

1. **Régression linéaire**
    - En utilisant la variable Average.Rooms, un modèle de régression linéaire simple est créé pour prédire la valeur médiane de la maison.
    - La fonction **lm** effectue la régression linéaire. La formule `Median.Home.Value ~ Average.Rooms` indique que Median.Home.Value est la variable dépendante, tandis que Average.Rooms est la variable indépendante.
    - La fonction **de résumé** est ensuite utilisée pour imprimer le résumé du modèle de régression linéaire, y compris les coefficients, la valeur R au carré, les valeurs p, etc.
2. Diagramme de dispersion de régression

- **Mise en place de l'intrigue**

```
regression_scatter <- ggplot(data, aes(x=Average.Rooms, y=Median.Home.Value))
```

Initialise le tracé à l'aide de la `ggplot()` fonction. Le bloc de données `data` est spécifié comme source des données et les mappages esthétiques (`aes()`) indiquent que l'axe des x représentera le `Average.Rooms` et l'axe des y représentera le `Median.Home.Value` .

- **Points de dispersion**

```
geom_point( alpha=0.6)
```

Cela ajoute des points de dispersion au tracé. L' `alpha=0.6` argument définit la transparence des points, les rendant quelque peu translucides et aidant à visualiser les points de données qui se chevauchent.

Ligne de régression linéaire

```
geom_smooth(method="lm", col="red", se=FALSE)
```

Cela ajoute une ligne de régression linéaire au tracé. Les arguments accomplissent les actions suivantes :

- `method="lm"` précise qu'un modèle linéaire (lm) doit être utilisé.
- `col="red"` définit la couleur de la ligne de régression sur rouge.
- `se=FALSE` indique que la zone ombrée (erreur standard) autour de la ligne de régression ne doit pas être affichée.

- **Étiquettes et titres**

labs(title="Regression Scatter Plot of Average Rooms vs Median Home Value", x="Average Rooms", y="Median Home Value")

Cela définit les étiquettes et le titre du graphique. Le titre est défini sur Diagramme de dispersion de régression du nombre moyen de pièces par rapport à la valeur médiane des maisons, et les étiquettes des axes x et y sont définies respectivement sur Nombre moyen de pièces et Valeur médiane des maisons.

- **Thème**

theme_minimal()

Cela applique un thème minimaliste à l'intrigue, ce qui offre un aspect propre et simple.

Résumé
Le code crée un nuage de points qui visualise la relation entre Average.Rooms et Median.Home.Value . Le tracé comprend également une ligne de régression linéaire rouge, indiquant la tendance de la relation entre ces deux variables. L'aspect général du tracé est clair et simple grâce à l'utilisation de la theme_minimal() fonction.

Résultats du modèle de régression linéaire
Un point crucial est que la valeur R au carré [0,76 (0,7587)] est très élevée. Cela signifie que la variable indépendante (Average.Rooms) explique 76 % de la variation de la variable dépendante, Median.Home.Value.
La valeur p pour le coefficient Average.Rooms est inférieure à 0,001 – 0,00058, ce qui indique que le coefficient est statistiquement significatif, c'est-à-dire qu'il existe une forte relation positive entre le nombre moyen de pièces d'une maison et sa valeur médiane. En d'autres termes, à mesure que le nombre de pièces d'une maison augmente, sa valeur médiane augmente également. Pour chaque pièce supplémentaire, la valeur médiane d'une maison devrait augmenter de 7 188,00.
Le coefficient d'interception est de 3,7904, ce qui est la valeur de la valeur médiane d'une maison lorsque le nombre de pièces est de 0. Ce scénario n'est pas réaliste mais aide à interpréter le coefficient de pente.
L'erreur standard résiduelle est de 2,699, ce qui correspond à la distance moyenne entre les valeurs médianes observées des maisons et les valeurs prédites par le modèle.
La statistique F est de 1855, ce qui est très élevé et indique que le modèle s'adapte bien aux données. La valeur p pour la statistique F est inférieure à 0,001, ce qui signifie que le modèle est statistiquement significatif.
Les résultats de la régression linéaire suggèrent une forte relation positive entre le nombre moyen de pièces d'une maison et sa valeur médiane. Le modèle s'adapte bien aux données et peut être utilisé pour prédire la valeur médiane d'une maison en fonction du nombre de pièces.

Notez certains aspects supplémentaires suivants à retenir lors de l'interprétation des résultats :

- Le modèle a été formé sur un ensemble de données de 489 observations, mais 15 ont été imputées en raison de données manquantes.
- Les résultats ne sont valables que pour la population de foyers représentée dans l'ensemble de données.
- D'autres facteurs (taille de la maison, emplacement, état) peuvent également affecter sa valeur médiane.

À partir du code ci-dessus, nous pouvons déduire que l'objectif principal est de comprendre la relation entre le nombre moyen de pièces dans les maisons et leurs valeurs médianes. L'histogramme fournit des informations sur la distribution des prix des maisons, tandis que l'analyse de régression linéaire décrit quantitativement la relation entre les deux variables.

Interprétation des données

Sur la base de notre analyse, nous pouvons voir que la taille de l'appartement détermine le prix d'une maison. Mais c'est rarement le seul facteur dans l'immobilier.

L'emplacement des transports en commun, des établissements d'enseignement et des maisons est également important. Nos données ont montré une faible corrélation entre l'accès aux transports en commun et les établissements d'enseignement. Mais cela pourrait être dû aux données elles-mêmes. L'accès aux transports en commun mesure le nombre de bus qui traversent le quartier en une heure. Mais, surtout, il ne mesure pas où les bus circulent. De plus, il ne fait pas la distinction entre les différents modes de transport, y compris les chemins de fer et les options terrestres comme les bus. Cette différenciation peut s'avérer vitale, en fonction des caractéristiques spécifiques de la ville.

L'accès à l'école publique est crucial pour l'attrait d'une maison lorsque des parents ou des couples emménagent dans une zone. Mais considérer la façon dont cette variable est mesurée n'est peut-être pas la meilleure façon de mesurer l'accès à l'école. Au lieu de simplement recenser le nombre d'écoles dans la région, il serait prudent de se concentrer sur les écoles exceptionnelles connues pour leurs impressionnants résultats aux examens des élèves. En outre, imaginez un scénario où l'abondance des écoles dans la région est due à sa densité de population. À mesure que la population augmente, les écoles deviennent surpeuplées et il ne reste plus de places disponibles. Notre base de données n'intègre cependant pas ce facteur crucial, malgré son impact considérable.

Le nombre d'écoles ouvertes n'est pas le seul facteur à prendre en compte ; la proportion d'espaces accessibles et la qualité de l'enseignement sont des facteurs importants qui peuvent influencer le choix d'une personne de déménager dans une zone spécifique. Ces facteurs peuvent en fin de compte avoir un impact sur les prix de l'immobilier.