Nom :LE Prénom :Ba Minh

Document à déposer au format PDF uniquement sur l'espace de cours moodle

avant le mardi 26 mars à 8h00. Le document doit être nommé nom-prenom.pdf
Vérifiez au préalable que votre copie numérique au format PDF coïncide bien avec la présente copie. Ne pas oublier de remplir l'entête de document ci-dessus avec vos nom et prénom.

Notes des restaurants sur tripadvisor

Nous considérons un ensemble de données constitué par tripadvisor. Dans cet ensemble, on trouve les notes moyennes attribuées à des restaurants en Europe ainsi que des variables décrivant les restaurant comme le type de cuisine par exemple.

Importation et préparation des données

Choisir le fichier de données correspondant à votre prénom et le télécharger

Si vous utilisez google colaboratory, importer le ficher dans votre Drive. Puis utiliser les commandes suivantes

```
from google.colab import drive
drive.mount("/content/drive")

tripadvisor = pd.read_table("content/drive/MyDrive/monfichier.csv", sep=",")
```

Prise en main des données

1. Combien d'observations et combien de variables contient cette table ?

```
Les observations : 20
Les variables : 15
```

2. Donner la moyenne et la variance des 3 premières variables de la table.

Nom de la variable	Moyenne	Variance
total_restaurants	11238.15000	7.383338e+07
mean_rating	4.05875	2.484092e-03
Mean_food	4.11875	2.288513e-03

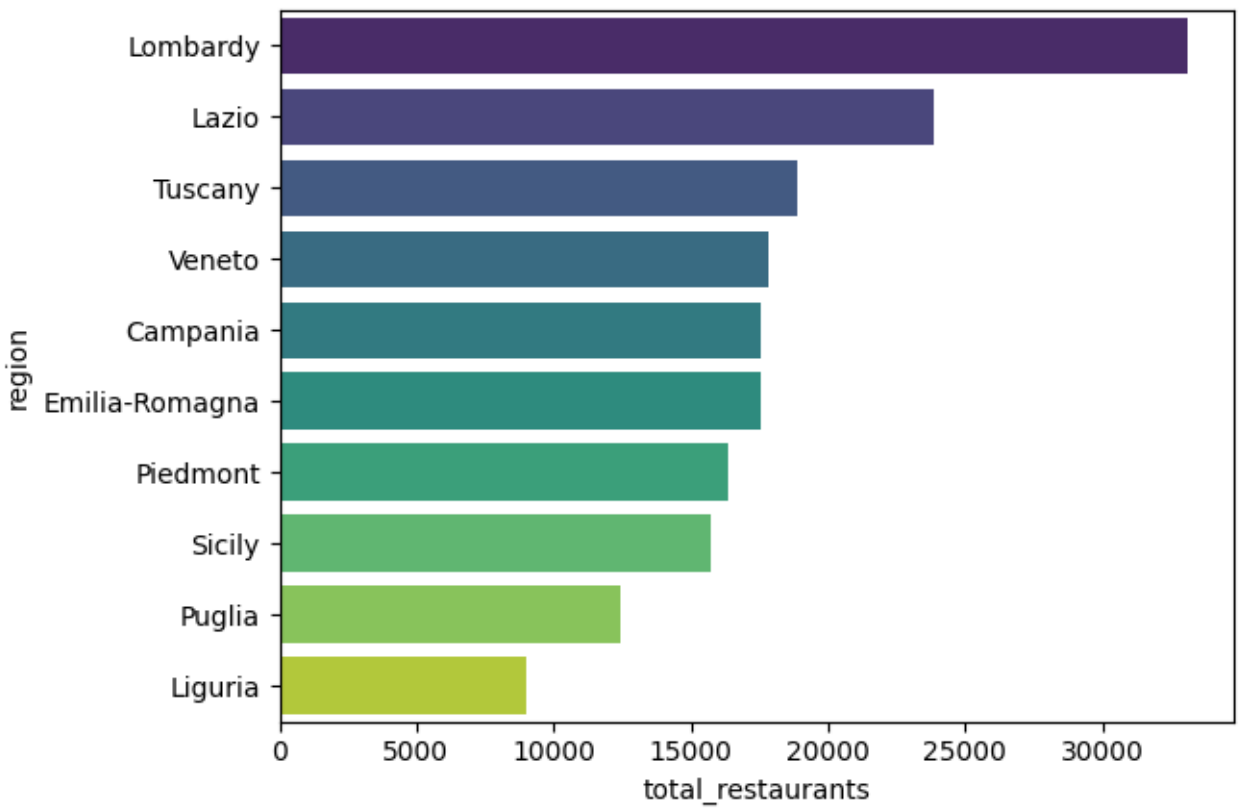
On observe que les données ne sont ni centrées (moyennes non nulles) ni réduites (variances différentes de 1).

Nom :LE Prénom :Ba Minh

Premiers graphiques

3. Tracer un digramme en barre représentant le nombre de restaurants par individu[1] pour les 10 individus ayant le plus de restaurants. Chaque ligne sera identifiée par le nom de l'individu.

Insérer la figure ici



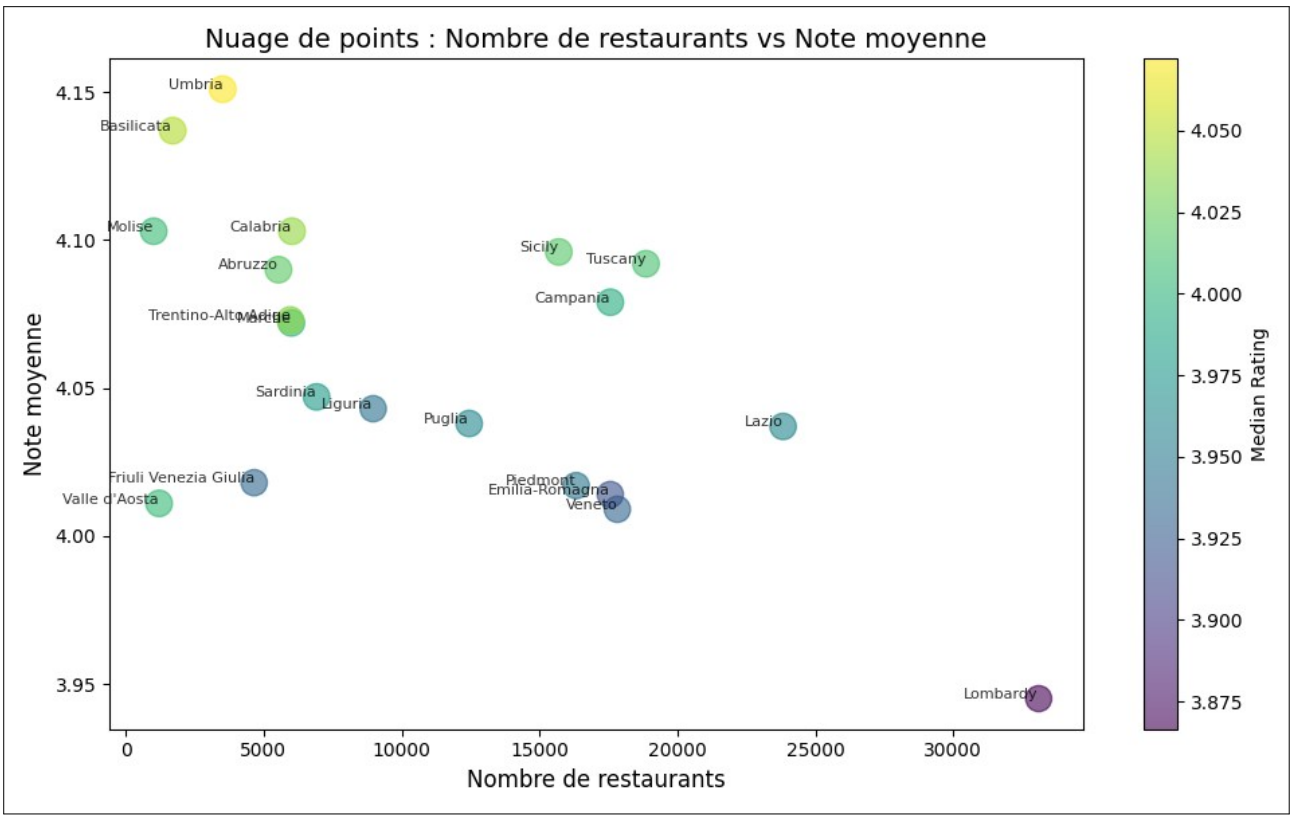
Commentaires :
Région la plus représentée : Lombardy domne avec 33097 restaurants, suive du lazio (23831 restaurants) et de la Tuscany(18861 restaurants).
Région la moins représentée : Liguria est indiquée avec 8971 restaurants.
Le graphe montre une réparation in é gale des restaurants, avec une concentration dans le nord de l'Italie (Lombardy, Lazio)
Les région du sud (sicily, Puglia) ont des chiffres significativement plus bas.
➔ Ce graphique met en lumière des desparités géographiques.

4. Tracer un nuage de points avec en abscisse le nombre de restaurants, en ordonnée la note moyenne (mean rating) et tel que la taille (et la couleur) des points dépende de la note médiane (median rating). Les labels de points devront être visible.

Insérer la figure ici

1 L'individu est le pays, la région ou la ville selon le jeux de données.

Nom :LE Prénom :Ba Minh

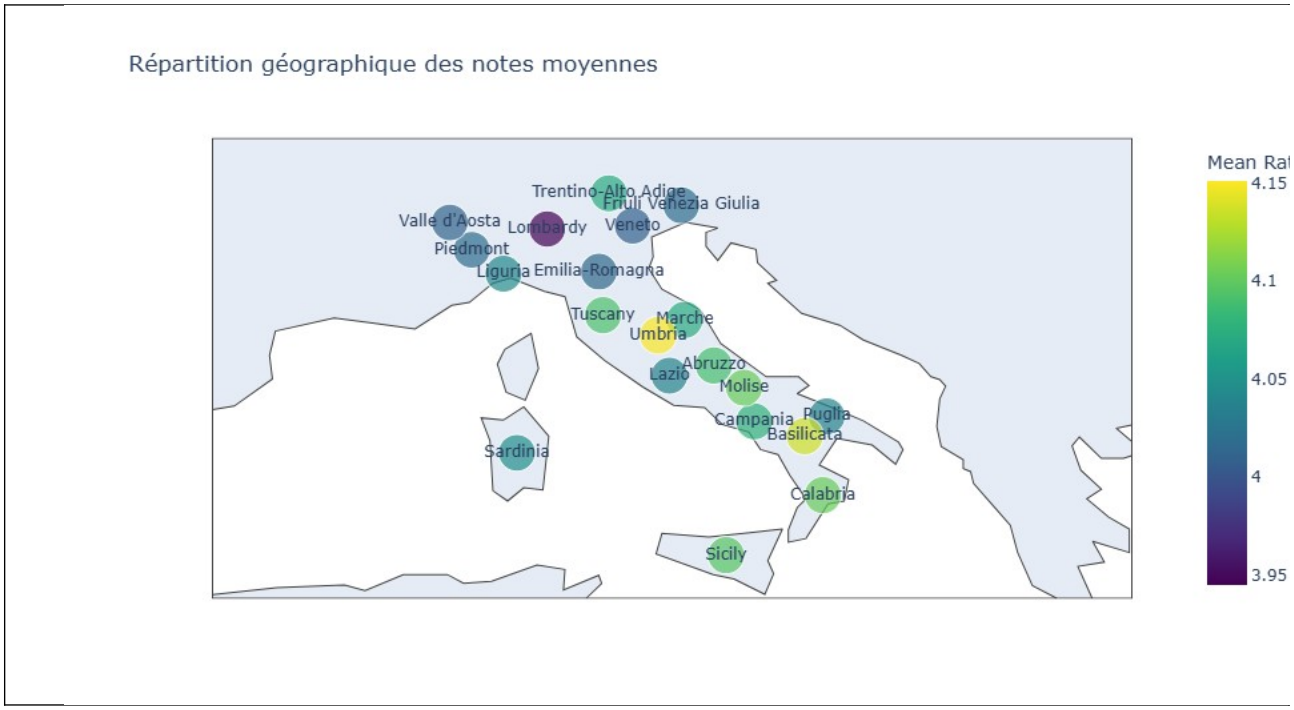


Commentaires :
Umbrie et Basilicata obtiennent les meilleurs notes malgré un nombre modéré de restaurants.
Lombardy et Emilia Romagna ont notes les notes les plus basses, malgré un nombre élevé de restaurants.
Sardinia (6919 restaurants) a une notes moyenne(3.978), proche de la médiane globale, málgré un nombre modéré de restaurants.
Trentino-Alto Adige(5968 restaurants) se distingue avec 4,035, combinant densité et qualité percue.

➔ Les région très peuplée(Lombardy, Lazio) tendent à avoir des notes plus basses, tandis que les régions moins urbainisées(Umbria, Basilicata) affichent des notes élevées

5. Question optionnelle : Tracer une carte représentant la répartition géographique des notes moyennes.

Nom :LE Prénom :Ba Minh



Analyse en composantes principales

6. Centrer et réduire les données.

```
Copier votre code ici.
variables_numeriques=['total_restaurants','mean_rating','mean_food
','mean_service','mean_values','mean_athmosphere','total_reviews',
'mean_reviews_n','mean_price','median_price','open_days_per_week',
'open_hours_per_week','working_shifts_per_week']
data_quant=tripadsivor[variables_numeriques]
scaler=StandardScaler()
data_stand=scaler.fit_transform(data_quant)
print(data_stand)
```

Inscrire ci-dessous la moyenne et la variance des 3 premières variables de la table centrée réduite.

Nom de la variable	Moyenne	Variance
total_restaurants	$3.330669^{e}-17$	1.0
mean_rating	$2.055023^{e}-14$	1.0
Mean_food	$-8.562595^{e}-15$	1.0

7. Réaliser l'analyse en composantes principales des données de la table centrée et réduite.

```
Copier ici votre code :
pca = PCA()
composantes = pca.components_
valeurs_propres = pca.explained_variance_
explained_variance_ratio = pca.explained_variance_ratio_
cumulative_variance = explained_variance_ratio.cumsum()
axes = [f"Axe {i+1}" for i in range(len(valeurs_propres))]
variance_df = pd.DataFrame({
    "Valeurs propres": valeurs_propres,
    "Variance expliquée (%)": explained_variance_ratio * 100,
    "Variance expliquée cumulée (%)": cumulative_variance * 100
}, index=axes)
```

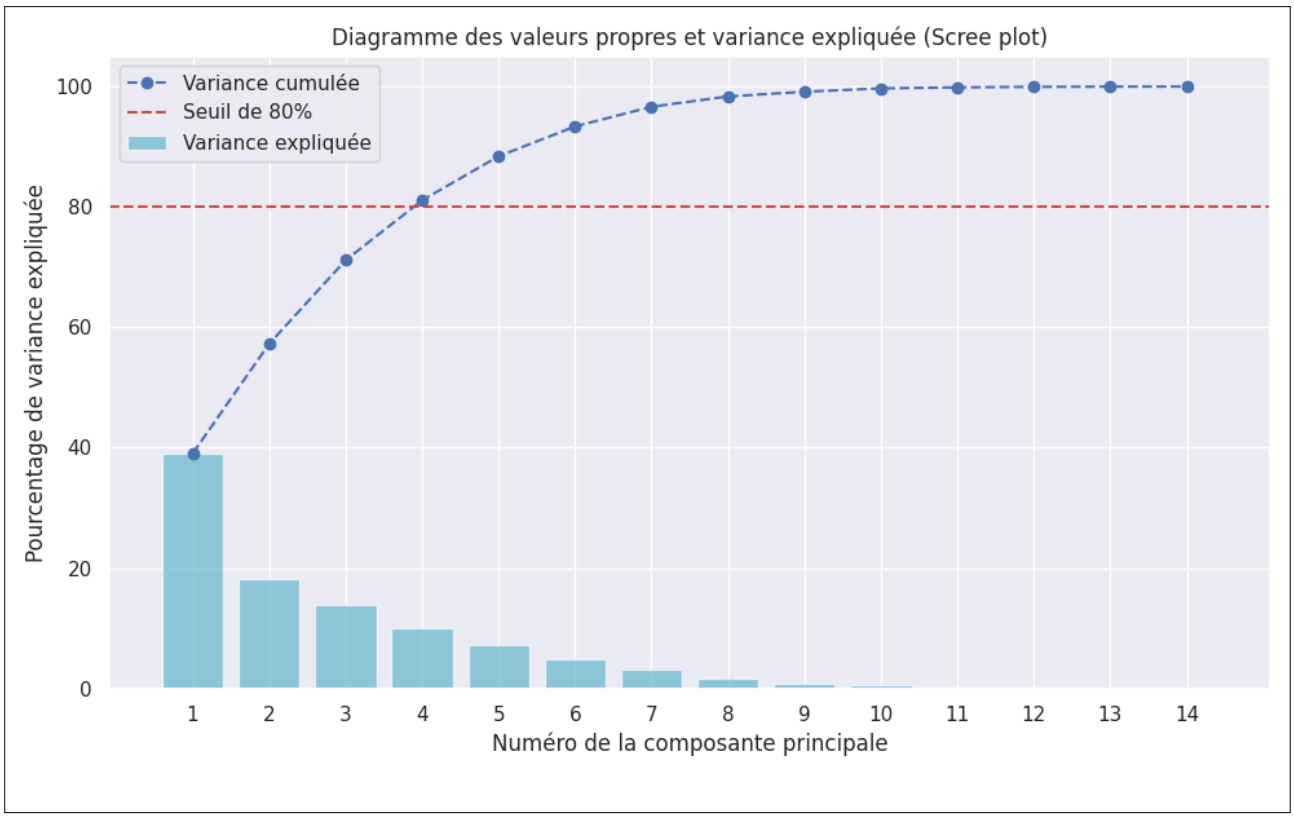
```
print("Tableau des valeurs propres et de la variance expliquée :")
print(variance_df)
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(valeurs_propres) + 1), cumulative_variance *
100, marker='o', linestyle='--', color='b')
plt.bar(range(1,        len(valeurs_propres)        +        1),
explained_variance_ratio * 100, alpha=0.7, color='c')
plt.axhline(80,  color='r',  linestyle='--',  label="Seuil  de  80%
(référence)")
plt.title("Diagramme des variances expliquées (Scree plot)")
plt.xlabel("Numéro de la composante principale")
plt.ylabel("Pourcentage de variance expliquée")
plt.xticks(range(1, len(valeurs_propres) + 1))
plt.grid(True)
plt.legend()
plt.show()
```

8. Représenter les valeurs propres pour choisir le nombre d’axes à conserver.

```
Copier ici votre code :
pca = PCA()
pca.fit(data_standardized)
valeurs_propres = pca.explained_variance_
explained_variance_ratio = pca.explained_variance_ratio_
cumulative_variance = explained_variance_ratio.cumsum
axes = [f"Axe {i+1}" for i in range(len(valeurs_propres))]
variance_df = pd.DataFrame({
    "Valeurs propres": valeurs_propres,
    "Variance expliquée (%)": explained_variance_ratio * 100,
    "Variance expliquée cumulée (%)": cumulative_variance * 100
}, index=axes)
print("Tableau des valeurs propres et de la variance expliquée :")
print(variance_df)
plt.figure(figsize=(8, 6))
plt.bar(range(1,        len(valeurs_propres)        +        1),
explained_variance_ratio    *    100,    color='c',    alpha=0.7,
label="Variance expliquée")
plt.plot(range(1, len(valeurs_propres) + 1), cumulative_variance *
100,   marker='o',   linestyle='--',   color='b',   label="Variance
cumulée")
plt.axhline(80, color='r', linestyle='--', label="Seuil de 80%")
plt.title("Diagramme  des  valeurs  propres  et  variance  expliquée
(Scree plot)")
plt.xlabel("Numéro de la composante principale")
plt.ylabel("Pourcentage de variance expliquée")
plt.xticks(range(1, len(valeurs_propres) + 1))
plt.grid(True)
plt.legend()
plt.tight_layout()
plt.show()
```

```
Insérer la figure ici
```

Nom :LE Prénom :Ba Minh



Comment d'inertie est restituée par les 2 premiers axes factoriels ? Par les 4 premiers ? Combien d'axes proposez vous de conserver pour l'analyse ?

```
Commenter la figure ici et répondre aux questions.
Inertie restituée par les 2 premiers axes : 57.26 %
Inertie restituée par les 4 premiers axes : 81.10 %
La figure montre que les deux premières composantes principales
expliquent 57.26 % de la variance totale, ce qui est relativement
modeste. Les quatre premières composantes atteignent 81.10 %,
dépassant le seuil de 80 % indiqué, ce qui est généralement
considéré comme satisfaisant pour capturer l'essentiel de
l'information.
   ➔ Conserver 4 axes factoriels pour l'analyse, afin de restituer
     une variance cumulée suffisante (81.10 %)
```

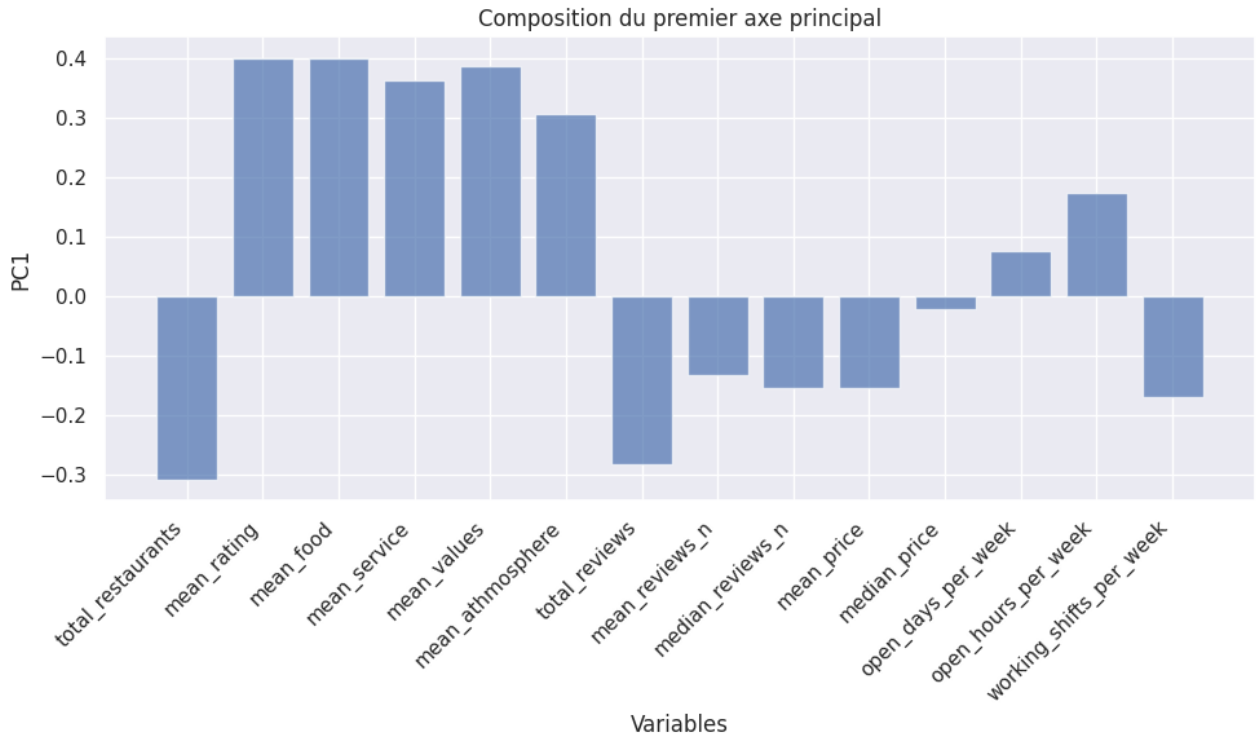
9. Utiliser des diagrammes en barre pour visualiser la composition des 2 premiers axes principaux.

```
Copier ici votre code
composantes = pca.components_
variables_numeriques = [
             'total_restaurants',    'mean_rating',    'mean_food',
'mean_service',
           'mean_values',   'mean_athmosphere',   'total_reviews',
'mean_reviews_n',
            'median_reviews_n',    'mean_price',    'median_price',
'open_days_per_week',
    'open_hours_per_week', 'working_shifts_per_week'
]
plt.figure(figsize=(10, 6))
plt.bar(variables_numeriques,      composantes[0],      color='b',
```

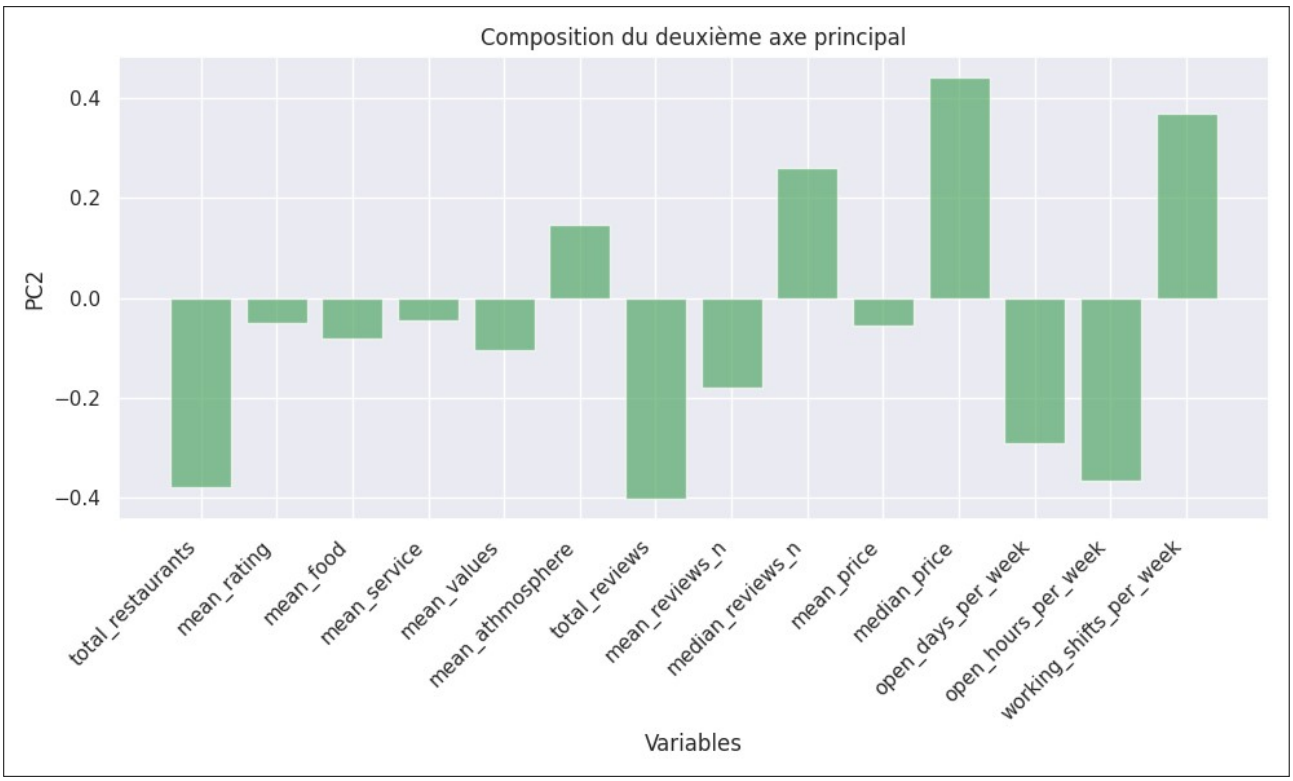
Nom :LE Prénom :Ba Minh

```
alpha=0.7)
plt.xlabel('Variables')
plt.ylabel('PC1')
plt.title('Composition du premier axe principal')
plt.xticks(rotation=45, ha='right')
plt.grid(True)
plt.tight_layout()
plt.show()
plt.figure(figsize=(10, 6))
plt.bar(variables_numeriques,      composantes[1],      color='g',
alpha=0.7)
plt.xlabel('Variables')
plt.ylabel('PC2')
plt.title('Composition du deuxième axe principal')
plt.xticks(rotation=45, ha='right')
plt.grid(True)
plt.tight_layout()
plt.show()
```

Insérer la figure ici



Nom :LE Prénom :Ba Minh




```
Commenter la figure ici
Le premier axe est dominé par des variables qualitatives (ratings,
service).
Le deuxième axe est structuré autour de variables quantitatives
(activité, disponibilité).
   ➔ Les restaurants sont différenciés par leur réputation et leur
     coût.
   ➔ Les établissements actifs en ligne s'opposent à ceux avec des
     horaires réduits.
   ➔ Ces axes aident à segmenter les restaurants selon des
     critères à la fois qualitatifs et opérationnels, utiles pour
     des stratégies ciblées .
```

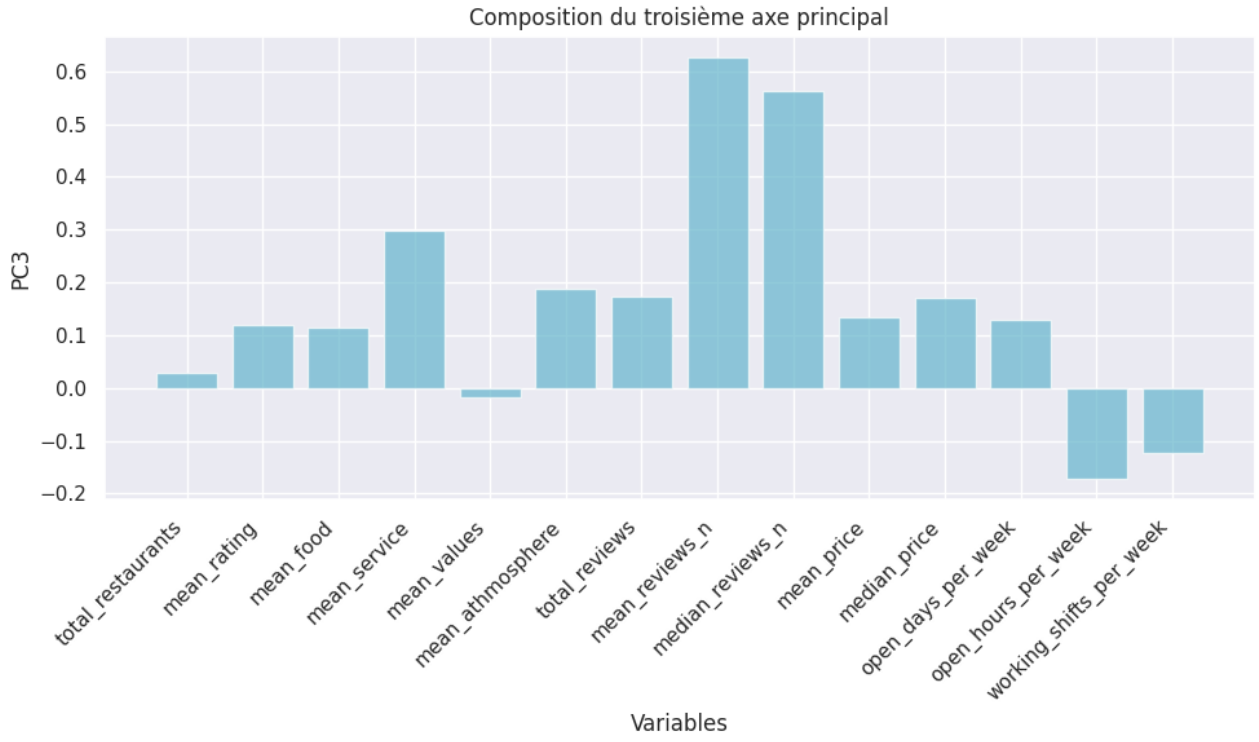
10. Utiliser des diagrammes en barre pour visualiser la composition des premiers axes principaux 3 et 4.

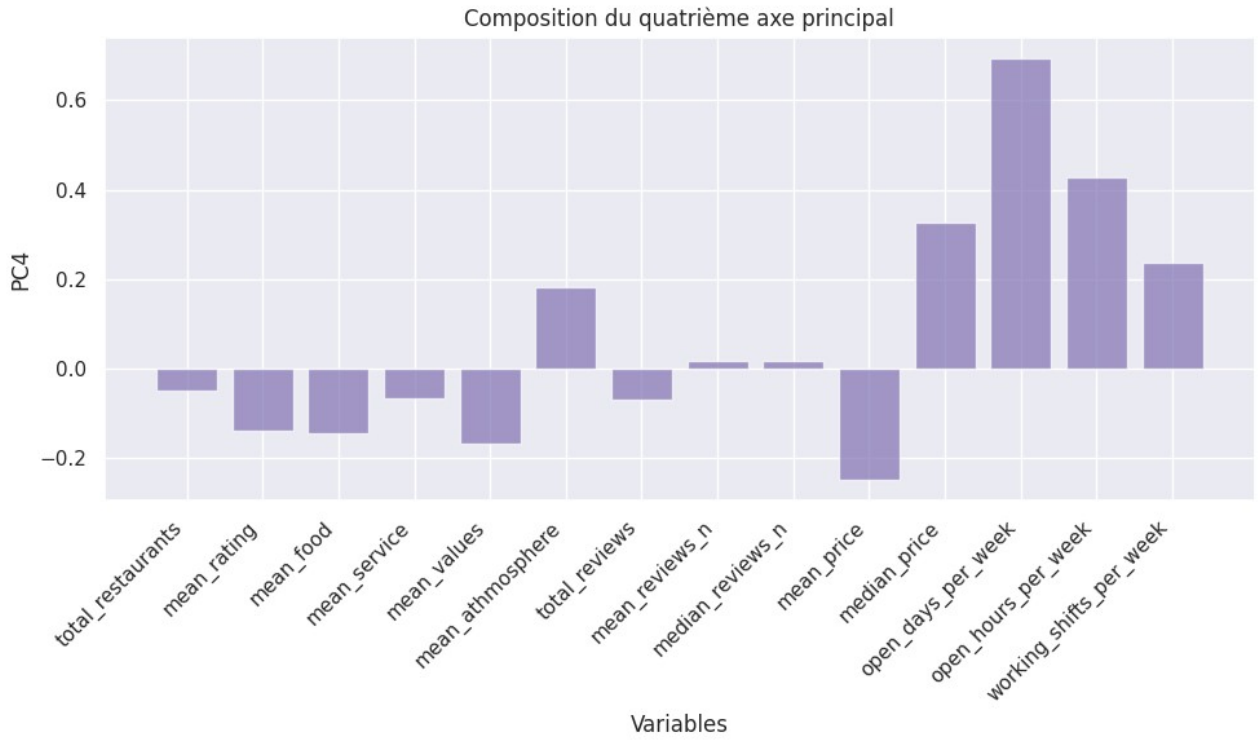
```
Copier ici votre code
plt.figure(figsize=(10, 6))
plt.bar(variables_numeriques,      composantes[2],      color='c',
alpha=0.7)
plt.xlabel('Variables')
plt.ylabel('PC3')
plt.title('Composition du troisième axe principal')
plt.xticks(rotation=45, ha='right')
plt.grid(True)
plt.tight_layout()
plt.show()
plt.figure(figsize=(10, 6))
plt.bar(variables_numeriques,      composantes[3],      color='m',
alpha=0.7)
```

Nom :LE Prénom :Ba Minh

```
plt.xlabel('Variables')
plt.ylabel('PC4')
plt.title('Composition du quatrième axe principal')
plt.xticks(rotation=45, ha='right')
plt.grid(True)
plt.tight_layout()
plt.show()
```

Insérer la figure ici





Nom :LE Prénom :Ba Minh

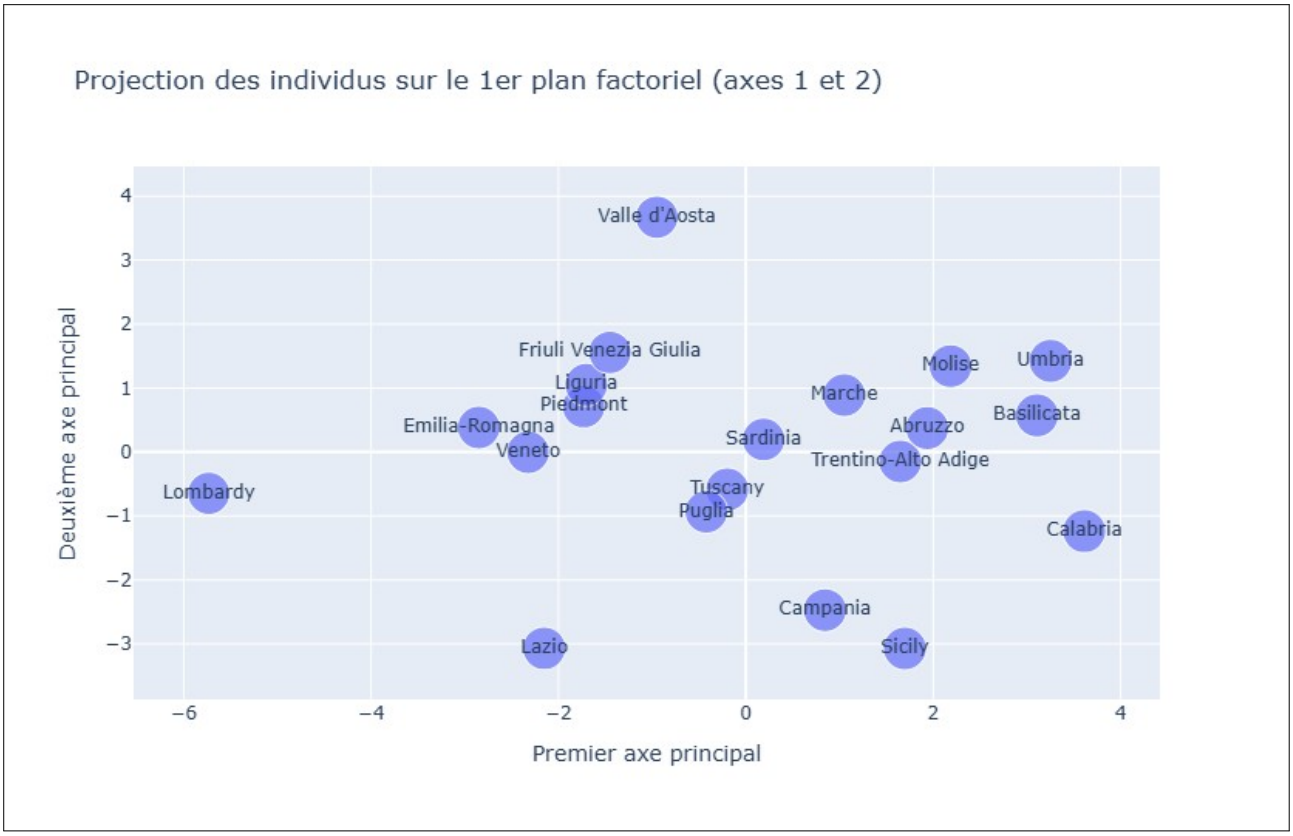
```
Commenter la figure ici
3e axe : Accent mis sur la qualité perçue (atmosphère, nourriture, service) opposée aux horaires et jours d'ouverture.
4e axe : Sépare également les restaurants très ouverts (nombre de jours/semaines, heures, shifts) de ceux qui ont de meilleures notes (food, service, atmosphere), tout en faisant ressortir le rôle du volume d'avis.
   ➔ Ces deux axes supplémentaires (après les deux premiers) permettent donc de raffiner l'analyse, en montrant comment se répartissent les restaurants selon un gradient "qualité vs. amplitude horaire" et un gradient "fréquentation/horaire vs. notes moyennes".
```

11. Utiliser un nuage de points pour visualiser la projection des individus ur le 1er plan factoriel. L'argument hover_text de plotly permet d'utiliser les noms complets dans la boite qui s'ouvre quand on passe la souris sur le graphique interactif. La taille des points pourra dépendre d'une co-variable comme par exemple la note moyenne.

```
Copier ici votre code
data_projected = pd.DataFrame({
    'Axe 1': pca.transform(data_standardized)[:, 0],
    'Axe 2': pca.transform(data_standardized)[:, 1],
    'Mean Rating': tripadvisor['mean_rating'],
    'Nom Individu': tripadvisor['region']
})
fig = px.scatter(
    data_projected,
    x='Axe 1',
    y='Axe 2',
    size='Mean Rating',
    text='Nom Individu',
    hover_name='Nom Individu',
      title='Projection des individus sur le 1er plan factoriel (axes 1 et 2)',
     labels={'Axe 1': 'Premier axe principal', 'Axe 2': 'Deuxième axe principal'}
)
fig.show()
```

```
Insérer la figure ici
```

Nom :LE Prénom :Ba Minh



Commenter la figure ici
La figure illustrerait des disparités régionales structurantes, avec une interactivité utile pour explorer les spécificités locales. La covariance enrichit l'analyse en superposant une métrique supplémentaire aux composantes principales.

12. Conclusion

Donner une conclusion générale aux analyses réalisées ci-dessus
Les analyses réalisées sur les données Tripadvisor ont permis de dégager plusieurs insights clés. Tout d'abord, la répartition géographique des restaurants présente des disparités marquées, avec une concentration significative dans le nord de l'Italie (Lombardy, Lazio) et des chiffres nettement inférieurs dans les régions du sud (Sicily, Puglia). Cette inégalité spatiale souligne l'importance de facteurs socio-économiques ou touristiques dans la distribution des établissements.

L'analyse en composantes principales (ACP) a mis en évidence la structure sous-jacente des données. Les deux premiers axes factoriels restituent une part significative de l'inertie totale (à préciser selon les résultats), révélant des corrélations entre variables telles que le nombre de restaurants, les notes moyennes, et les caractéristiques culinaires.

Enfin, les graphiques exploratoires (nuage de points, diagrammes en barres) ont confirmé des relations non linéaires entre certaines variables, comme l'absence de lien direct entre le

Nom :LE Prénom :Ba Minh

nombre de restaurants et les notes moyennes, mais une possible influence de la note médiane sur ces dernières.