Projet SAS

# Partie 1 : IMPORT ET TRAITEMENT DES DONNÉES

Tout d’abord nous avons importé le fichier *Sashelp.Demographics* dans notre propre bibliothèque afin d’y appliquer des modifications dans le nom de certains pays. Ces modifications ont été appliqués afin de pouvoir fusionner les données de la table *Demographics* et *Pays\_covid\_hebdo\_1*, selon la colonne *Pays* d’où la nécessité de devoir modifier les noms de certains pays dans *Demographics* pour que la fusion soit bien appliqué.

Après quelques manipulations sur SAS, nous nous retrouvons avec les deux tables sur lesquelles vont se poser notre analyse; *PAYS\_covid\_act* et *IDX\_covid\_hebdo*.

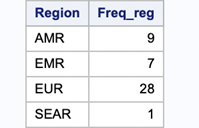
# Partie 2 : ÉLABORATION DE DONNÉES STATISTIQUES

a)

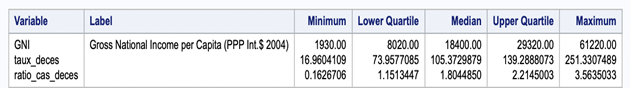
Afin de proposer des analyses des pays les plus touchés en termes de nombre d’infections et de décès, on a décidé de diviser ces données par la population totale de chaque pays, puis de multiplier par 100 000 afin d’obtenir un ratio sur 100 000 habitants. Les colonnes nouvellement créées sont nommées taux\_deces et taux\_cas.

La manière dont nous nous sommes pris pour déterminer les pays les plus touchés en termes d’infections et de décès a été d’abord d’obtenir le troisième quartile des deux taux puis de sélectionner les observations ayant un taux plus élevé que le troisième quartile obtenu. À partir des données obtenues, on a déterminé la fréquence des régions ayant un pays avec un taux de cas et de décès supérieur au 3ème quartile ainsi que des statistiques descriptives.

*Tableau 1: Fréquence des régions présentant des pays avec un taux d'incidence élevé*

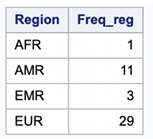


*Tableau 2: Statistiques descriptives des pays présentant un taux d'incidence élevée*



On observe que les pays présentant un taux d’incidence élevée se trouve en grande majorité en Europe. Ensuite, on a sorti les statistiques descriptives de ces pays selon le GNI le taux de décès et le ration cas/décès. Au tableau 2, on remarque que les pays plus touchés sont relativement riches en général et présente un taux de décès tout aussi élevé. En effet, le 3ème quartile des taux de décès est de 72.706663, alors que le quartile inférieur des pays les plus touché est de 73.9577085, démontrant une certaine corrélation entre cas et décès et cela malgré la richesse et ressources en possessions d’un pays.

*Tableau 3: Fréquence des régions démontrant le plus pays avec un taux de décès supérieur au 3ème quartile*

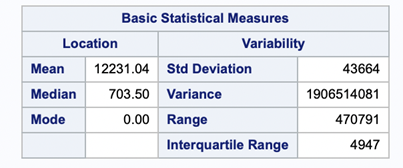


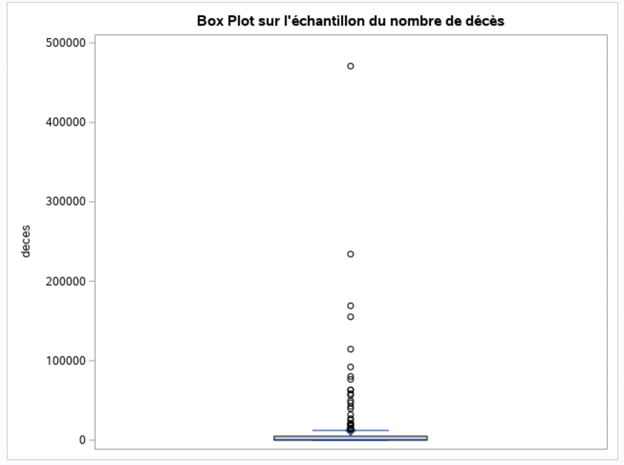
*Tableau 4: Statistiques desccriptives des pays présentant un taux de décès élevé*



Comme pour les pays démontrant un haut taux d’infection, les pays avec un taux de décès élevé se trouve majoritairement en Europe. Le tableau 4 confirme les remarques apportées pour le taux d’incidence.

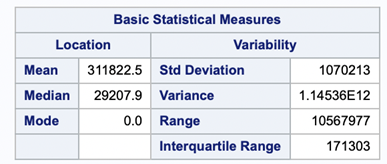
*Tableau 5: Indicateur de position et de dispersion pour la variable décès*

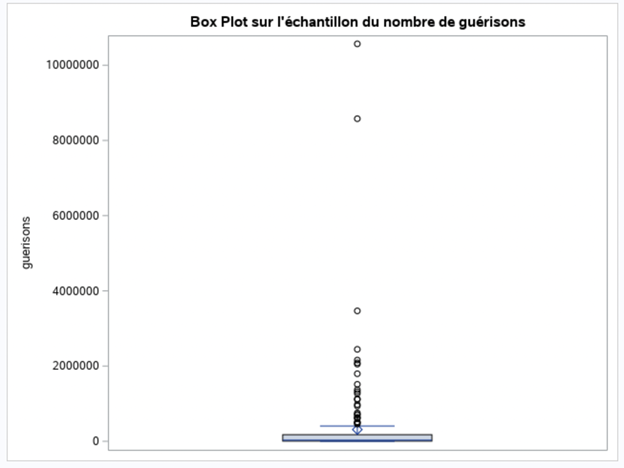




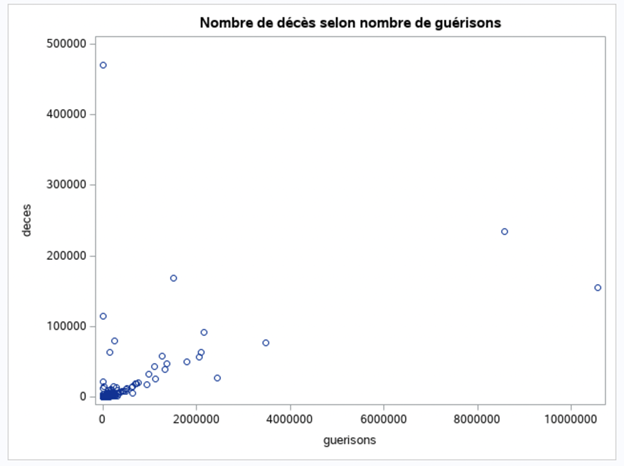
*Figure 1: Boxplot de la variable décès*

*Tableau 6: Indicateur de position et de dispersion pour la variable guérison*





*Figure 2: Boxplot de la variable guérison*



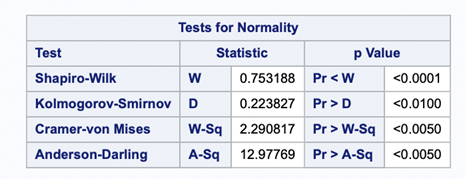
*Figure 3: Nuage de points du nombre de décès et de guérisons par pays*

À partir des tableaux et figures présentés ci-dessus, les variables *décès* et *guérisons* sont très variables. Cela est surtout dû aux observations, quelques peu nombreuses, aux extremums supérieurs. Ceci est clairement identifiables dans la figure 3, démontrant la majorité des observations avec un nombre de décès et de cas relativement faible aux quelques dizaines démontrant des valeurs importantes.

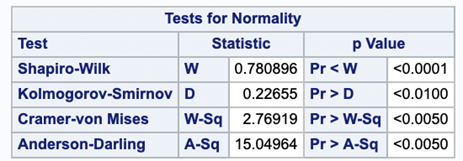
c)

Avant de tester la normalité des variables taux décès et taux d’infection, nous avons standardiser ces derniers afin que les données puissent être comparable entre elles. Pour cela, nous avons décidé d’utiliser la médiane comme valeur de référence et l'intervalle interquartile (divisé par deux) comme paramètre de dispersion. En effet, étant donné qu’on teste la normalité d’un échantillon et que nous ne connaissons pas sa distribution, standardiser selon la moyenne et l’écart-type n’est pleinement valable que lorsqu’elle appliqué à des distributions unimodales et symétriques. Ainsi, comme la distribution ne se montre pas symétrique et unimodale, il est préférable de standardiser selon la médiane et l’intervalle interquartile.

*Tableau 7: Test de Normalité pour le taux de cas standardisé*



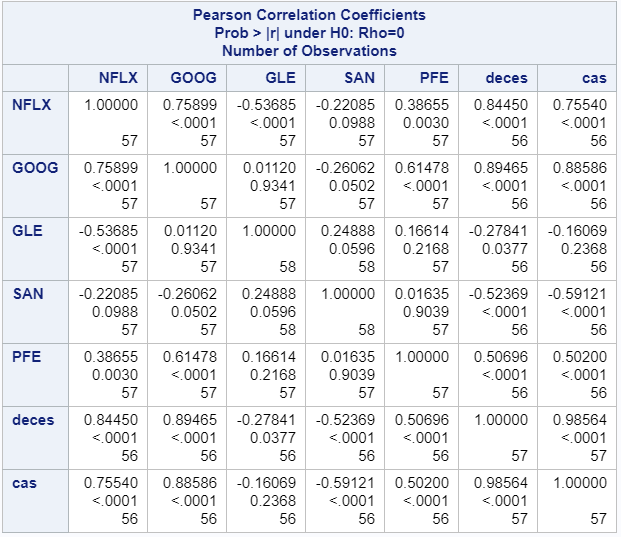
*Tableau 8: Test de Normalité pour le taux de décès standardiser*



À partir des tableaux 7 et 8, nous pouvons rejeter l’hypothèse nulle selon laquelle les variables analysées suivent une distribution normale.

# Partie 3 : ÉLABORATION DE MODÈLES LINEAIRES AVEC LA TABLE COVID.IDX.HEBDO:

4) Matrices de corrélation



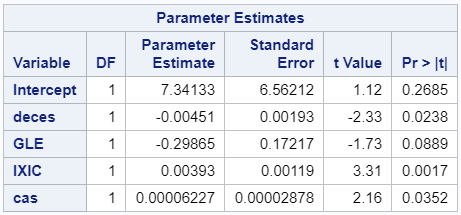
*Figure 4: Matrices de corrélation, coefficients*

L’indice Netflix a une corrélation fortement positive avec le nombre de décès et de cas et une corrélation négative modeste avec l’indice de la société générale. Il est possible qu'avec la progression de la pandémie les gens sont restées plus chez eux et du coup ils ont eu plus de temps pour regarder des séries, surtout avec la fermeture des cinémas. L’indice GLE de la société générale fait partie du CAC40 est peut reflète l’optimisme du marché alors que l’indice Netflix a plus tendance à augmenter avec les vagues plus fortes de l’épidémie. A notez aussi que l’indice Pfizer a des taux de corrélation avec le taux de cas et décès qui est plus bas que l’indice Netflix.

**5)**

**Model 1:** PFE = β0 + β1\*deces + β2\*GLE+ β3\*IXIC + β4\*cas + µ

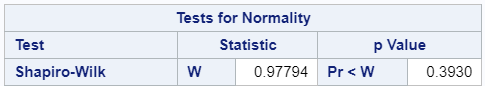
*Tableau 9: Coefficients de la régression*





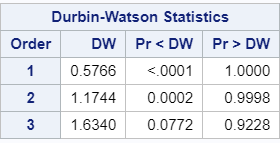
La régression du tableau 9 montre qu'une augmentation du taux de décès moyen mondial par semaine de un réduit en moyenne l’indice de Pfizer de 0.45 % . L’indice IXIC a un effet positif, montrent que l’indice Pfizer suit le marché. Lorsque la situation de Covid s’aggrave dans le monde les marches ont une tendance de baisser. Il est possible que le développement d’un vaccin puisse être un de ses facteur qui influence le marché positivement. Par contre, il faut interpréter ces coefficients avec prudence car on n'a pas vérifié les hypothèses du modèle linéaire

*Tableau 10: Test de normalité pour les résidus*



Pour le modèle 1, la normalité des résidus est conserve avec le test shapiro-wilk qui donnent un p-value de 0.39. On conserve donc l’hypothèse de H0 que les résidus sont distribués en fonction d’une loi normale. Ceci peut aussi être vérifier avec le plot “Normal - QQ” dans la proc reg.

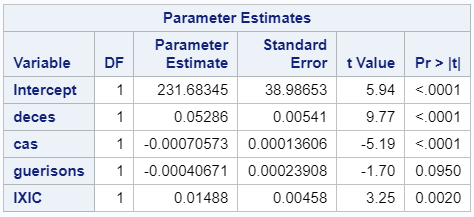
*Tableau 11: Test d'autocorrélation*



En revanche l’hypothèse d'Independence entre les observations successives n’est pas conservée. Cela est confirmé à l’aide d’une proc autoreg pour effectuer un test Durbin-Watson. On observe une autocorrélation positive au deuxième dégrée.

**Model 2:** NFLX = β0 + β1\*décès + β2\*cas+ β3\*guérisons+ β4\*IXIC+ µ

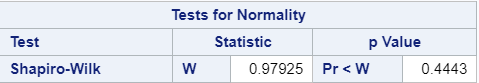
*Tableau 12: Coefficients de la régression 2*





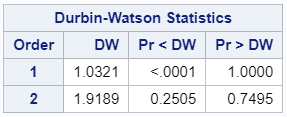
La régression du tableau 12 montre qu’une augmentation du taux moyen mondial de décès par un dans une semaine, augmentera l’indice de Netflix de 5% en moyenne. L’effet du marché est aussi plus fort pour l’indice Netflix compare à l’indice Pfizer, un résultat plutôt surprenant. Cependant on suppose encore une fois que les hypothèses de OLS tiennent, alors qu’il est clair que cette régression qui se base aussi sur un stock souffre de problème d’autocorrélation.

*Tableau 13: Test normalité des résidus du modèle 2*



Pour le modèle 2, la normalité des résidus est aussi conservée avec un p-value de 0.4443 vue dans le tableau 13.

*Tableau 14: Test d'autocorrélation pour le modèle 2*

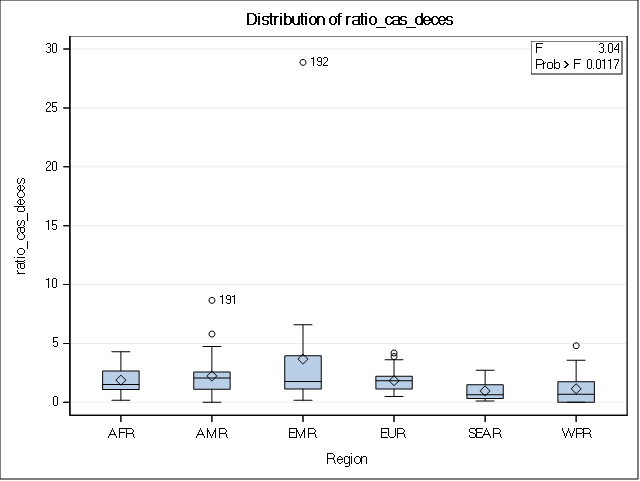


Cependant il n'y a pas l'Independence entre observations successives. La proc autoreg est une fois de plus utilisé pour confirme à l’aide d’un test Durbin-Watson vue dans le tableau 14. On conclut qu'il y a de l’autocorrélation positive au premier dégrée.

Partie 4 : DÈTERMINATION DE L’INFLUENCE DES CONTINENTS SUR LA PANDEMIE

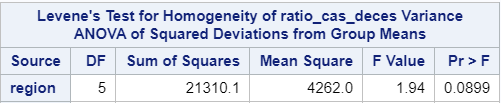
6)

**Modèle ANOVA 1 :** Ratio\_cas\_deces = Region



*Figure 5: Boxplot de la Proc ANOVA*

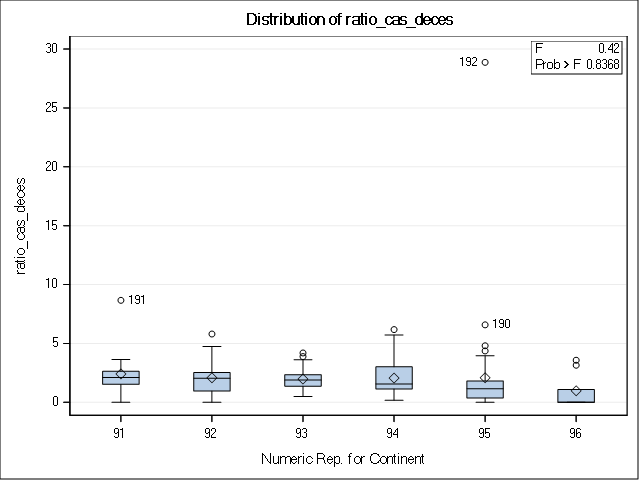
La variable région est utilisée comme variable catégorielle. Le boxplot montre le taux de l’létalité par région. On voit que l’EMR et la région avec le taux de létalité le plus élevé. Les “outliers” 191 et 192 sont le Yemen et le Mexique notée par leurs numéros d’observations après la proc sort. Avec la p-value = 0.01 du F-test l’hypothèse qu'il n'y a pas différence entre les moyennes d’au moins une des régions est aussi rejeté.



*Tableau 15: Test d'egalite des variances pour le modèle ANOVA 1*

A l’aide d’une proc univariate on effectue un test de normalité sur ratio cas décès par groupe. On conclut à l’aide de multiple test Shapiro-Wilk que les groupes ne suivent pas une distribution normale, (on a aussi vérifié ceci pour la variable ratio\_cas\_deces). On utilise aussi le teste Levene dans le tableau 15 qui donne un p-value de 0.09, on conclut que l’hypothèse H0 est accepté est que la variance entre les groupes sont similaires.

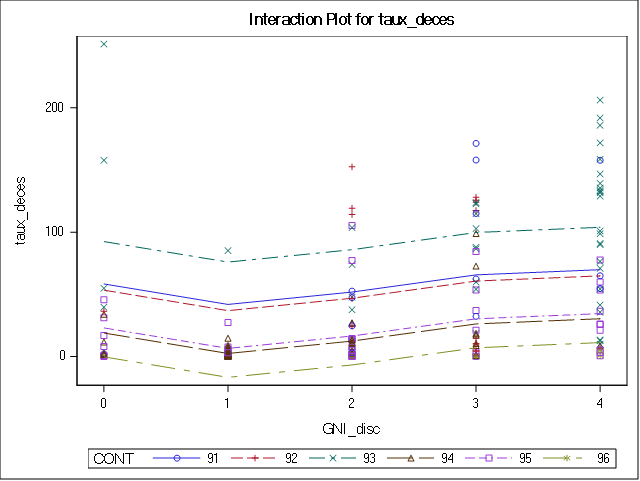
**Modèle ANOVA 2 :** Ratio\_cas\_deces = CONT



*Figure 6: Boxplot de la Proc ANOVA 2*

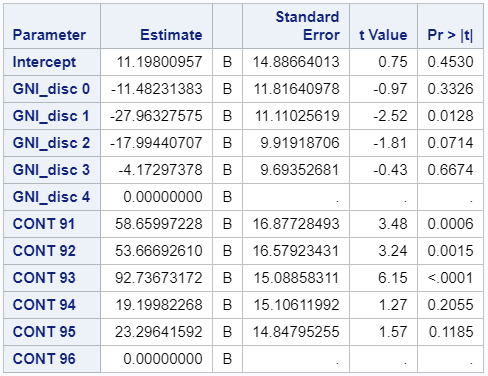
La Proc ANOVA avec la variable catégorielle continent donne avec le F-test un p-value de 0,83. Il a donc aucune différence de moyennes de taux de létalité entre les continents, un résultat différent comparer aux régions. Les tests Shapiro-Wilk et Levene sont aussi de nouveau utilisée. On conclut de nouveau que la variance entre les groupes est similaire mais que les données des groupes ne suivent pas une distribution normale.

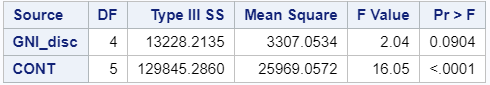
7) **Modèle GLM** : Taux\_deces\_std = GNI\_disc + Continent



*Figure 7: Résultat du Proc GLM, GNI\_disc de 0 = valeur manquante, GNI\_disc de 4 = Pays dans le top 25% des GNI*

*Tableau 16: Coefficients du modèle GLM*



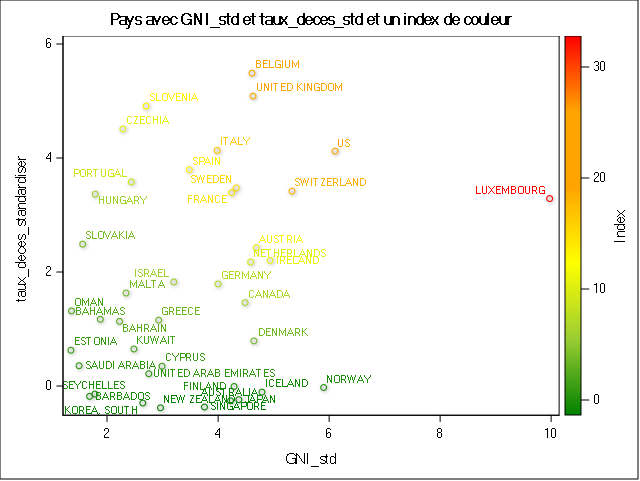


GNI\_disc = 0 sont des observations manquantes. La proc GLM décrit au-dessus corresponds à un modelé ANOVA a deux facteurs. Le F-test au-dessus conclut que seulement le facteur CONT est significatif après avoir contrôle pour le facteur GNI\_disc. Ce modèle repose sur les mêmes hypothèses décrit dans les modèles ANOVA au début de la Partie 4. On repère que les groupes n’ont pas le même numéro d’observation et que l’hypothèse de variance égale ne tient pas. (On avait aussi testé la normalité de taux\_deces dans la partie 2)

Figure 7 montre que les pays avec les GNI les plus élevé ont l’air d’être plus fortement touché en termes de décès. En plus le continent 93 (Europe) est celui qui a été le plus touché, alors que le continent 96 (L’Océanie) et le moins touché. Cela donne beaucoup de sens en termes de densité de population et de proximité d’autre pays. Le tableau 16 liste les coefficients de chaque variable “dummy” des deux facteur. Les deux catégories de référence sont L’Océanie et le GNI\_disc 4. Les autres “dummy” indique de combien le taux de cas changera compare à la catégorie de référence. Le P-value indique quels coefficients ont un effet significatif.

Partie 5 : Conclusion

8)



*Figure 8: Scatterplot avec taux de deces et GNI standardiser, l’index represente l’interaction de ces deux variables*

On estime que le taux\_deces reflète le risque de décès d'une meilleure façon que le ratio\_cas\_deces. On fait ce choix parce qu’on estime que là capacité de détecter les cas dépends de la capacité de tester et de tracer, ce qui fausse le taux de l’létalité (Rajgor et al., 2020).

On aperçoit dans la figure 8 a l’aide de l’index qui est compose d’une interaction entre les deux variable standardiser que les 5 pays avec le risque de décès le plus élevé et le revenu moyen le plus élevé sont: **Luxembourg, US, Belgium, UK, Switzerland**

Bibliographie

Rajgor, D. D., Lee, M. H., Archuleta, S., Bagdasarian, N., & Quek, S. C. (2020). The many estimates of the COVID-19 case fatality rate [Abstract]. *The Lancet Infectious Diseases,* *20*(7), 776-777. doi:10.1016/s1473-3099(20)30244-9