Student name

Course

Instructor

Format: Harvard

Date

Word count

# Introduction

Over the recent year, the internet has facilitated almost all dimensions of life. According to Isaac, Abdullah, Ramayah and Mutahar (2018), internet usage has been adopted in all sectors of the economy. Some of such sectors include; shopping, health informatics, education, banking, and transportation, among others.

The importance associated with access to the internet leads us to notice the widening gap between the first-world class and the third (developing) world-class countries. According to Yang and Zhang (2019), the uneven adoption of information communication technologies is expected to widen the gap between the two countries.

The unfortunate situation is that the developed countries are currently pressing on developing internet access while some developing countries are yet to test the fruits of the technology. As a result, globally, there is some information communication Technology (ICT) challenges that will still service unless poorer countries are rendered some assistance and catch up with developed countries (Matthess, 2020).

As a result, the current study seeks to develop a predictive model to predict the percentage of internet users based on the number of secured internet servers per 100 persons and the mobile cellular subscriptions per 100 people. The study data were retrieved online via the link https://www.worldbank.org/en/topic/digitaldevelopment.

# Literature review

The diffusion of innovation theory cited by Min and Jeong (2021) states that communication spreads from the point of origin through the surroundings. The theory continues to urge that when a small portion of the individual gets the invocation, it grows

According to Talebian and Mishra (2018), diffusion of innovation is comparable to the extent and the speed at which the economy can adopt new inventions. Therefore, the main concern is the flexibility of the economy to diffuse into new technologies, which can be immediate or slowly.

Empirical studies show that the rate at which innovation gets adopted in the economy depends on the innovators and the earlier adopters. Their efforts determine whether the technology will peak or not. However, internet development presents a unique opportunity to revise the innovation development theory. Mattsson and Andersson (2019) presented a unique argument from previous innovations. They add that the internet adapts areas such as animation and interactivity beyond the originally intended areas.

The internet push has as well been felt in business pull. For instance, in the study by lee et al. (2018), internet innovation has resulted in the business pull as organizations strive to have and maintain their online presence. As a result, several factors can be associated with internet use in the economy. For instance, a study conducted by Ishfaq and Mengxing (2021) associates internet usage with infrastructures, policies posed by the government and regulations. Other factors include; culture, penetration of the information technologies, and language. David and Grobler (2020) found that the internet usage rate would increase with the rate at which the economy of the given country grows. Previous studies have shown that higher income-generating groups tend to use the internet more than the low earning groups.

However, little attention has been focused on the impact of the number of secure internet servers and mobile cellular subscribers on internet usage (Bayar, 2021). Thus, this forms the main focus and objective of the current study.

# Methodology and data

The current study focused on the initial analysis of the existing data. The data were obtained from the World Bank database via the link <https://www.worldbank.org/en/topic/digitaldevelopment>. Three datasets were retrieved from the website. These include mobile cellular subscriptions (per 100 people), secure internet servers, and the individuals using the internet (percentage of the population).

A systematic data analysis methodology was employed. Where RStudio was employed as the platform for the analysis. First, after the three datasets were downloaded into the local computer, the preliminaries on RStudio were conducted. The preliminaries involved cleaning the memory, setting the working directory and loading the required libraries. This was then followed by loading the data into the program.

Given that all the three data were in long format. The reshpa2 package in r was used to shrink the data into a small format. This was done; for instance, considering the server dataset, the shrinkage was done such that country, date and servers were left as the only names of the columns. The next step was then to merge the data. The data merging was done using the country name and date as the unique identifiers. The final data has five columns. The country name, date, the number of secure servers, cellular mobile subscription and the percentage of the population using the internet.

A simple random sample of ten countries was selected from the data. The selected countries were in the top ten in terms of number. Of secure servers. This was followed by data preprocessing. In the data preprocessing, the first step was to remove the symbol X in the year column and then replace all the missing values with 0 in the data. This ensured that the data was good enough for the subsequent analysis.

In the data analysis, both descriptive-analytic and predictive analytics were employed. As part of the descriptive analysis, the data summary was computed. These seek to give a general view of the data. Besides, three-time plots were constructeucted. The plots explain the variation of the secure server, cellular subscribers and the percentage of internet users over time (Maleki, 2020). Furthermore, the total number of features was also compared across the countries using bar plots. Finally, k-mean clustering was conducted to try and determine if there is natural clustering of the data.

The ordinally least square (OLS) method was used for the prediction. OLS is a statistical method that seeks to determine the prediction equation of the response variable based on the explanatory variable(s) (Embaye, 2021). The model is conducted under the assumption that; there is a linear association between the response variable and the explanatory variable(s). it is also assumed that there is no or litter multicollinearity among the explanatory variables (Flatt, 2019). Besides, the residuals are assumed to be normality distributed with constant variance (Bun, 2019).

In this case, the percentage of the population using the internet was considered the response variable. On the other hand, the secure server and the cellular subscribers were the explanatory variables. After the model was conducted, the assumption was graphically assessed to determine the performance of the fitted model.

# Analysis and Data

## Descriptive summary

Three datasets were considered for this analysis. After being converted into a short format, each of the three data has a total of 16492 observations and three variables. We merged the three data into one with a total of 16492 observations and five variables. Among them includes country name, year, the percentage of users, number of secure internet server and mobile cellular subscriptions per 100 people. To continue with the analysis, the top ten countries interns of the percentage of internet users were selected for the analysis. The study focus study period from 1960 to 2021.

The selected countries were; the British Virgin Islands, Denmark, Seychelles, Netherlands, Singapore, United States, North America, Switzerland, Ireland, and Iceland. A total of 620 observations were then sampled. On average, the internet users' percentage was 23.83 %, ranging between 0 and 99.01 per cent. The number of secure servers ranged between 0 and 741079, with an average mean of 9284. The average mobile cellular subscription was found to range between 0 and 198.1522, where the mean was 37.67 subscriptions per 100 people, as illustrated in table 1.

Table 1 The summary statistics for the individual using the internet, the secure server and the mobile cellular subscriptions.

|  |  |  |  |
| --- | --- | --- | --- |
| statistics | Users (%) | Servers | Subscriptions (per 100 people) |
| Minimum | 0.00 | 0.00 | 0.00 |
| First quantile | 0.00 | 0.00 | 0.00 |
| Median | 0.00 | 0.00 | 0.38 |
| Mean | 23.82 | 9284.00 | 37.67 |
| 3rd quantile | 53.94 | 0.00 | 87.68 |
| maximum | 99.01 | 741079.00 | 198.15 |

Further, the ten countries were compared regarding the total percentage of individuals using the internet. Figure 1 shows that most internet users came from Iceland, 32.3%, followed by Denmark at 29.8%. Besides, internet usages were low in the British Virgin Islands, where the overall percentage of the population was 5.20%.

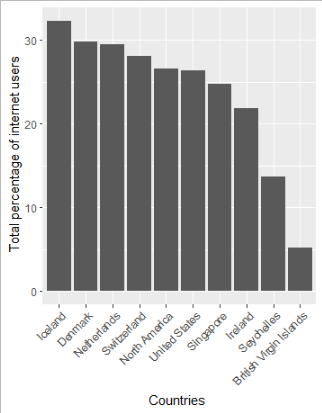


Figure 1 A comparison of the percentage of the population that uses the internet

Considering the number of secure internet servers, the majority (n=26656) were from the British Virgin Islands, whereas Iceland had the minority of secure servers (n=5393); *this is shown in figure 2.*

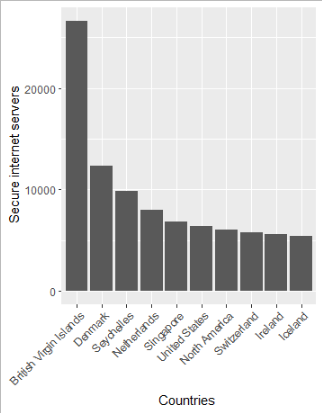


Figure 2 A comparison of the number of secure internet servers

Figure 3 shows that Singapore was the best country in mobile cellular subscriptions at 45 per 100 people. Besides, North America registered a minor subscription of 31 per 100 people.

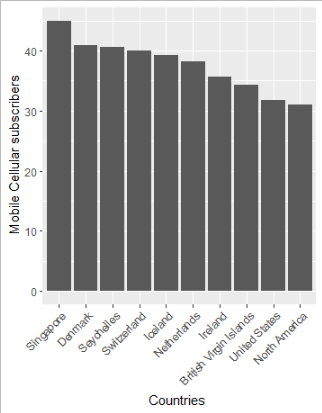


Figure 3 The mobile cellular subscription per 100 people

Further, the study seeks to explore the time series of the tree variables. The results in Figure four show a significant increase in all the variables just after 1990. This marked the eve of the 21st century when digital technologies marked significant innovation. It was observed that in the 21st century, digital literacy and technologies have advanced, which can be associated with the increased cellular mobile subscription, a high percentage of the population with an online presence and an increased number of secure servers.

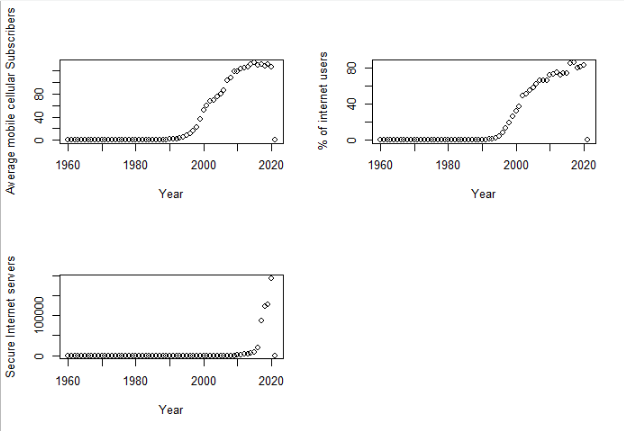


Figure 4 Time series plots for the percentage of internet users, number of secure servers and the mobile cellular subscription

## k-means clustering

further, the study seeks to determine if the data for the ten countries can be distinguished into k clusters. To achieve this, a k-means clustering algorithm was conducted. A scree plot was plotted to determine the optimal number of k, as shown in figure 5. Plot 3 was chosen as the best k. thus 3-mean clustering was conducted.

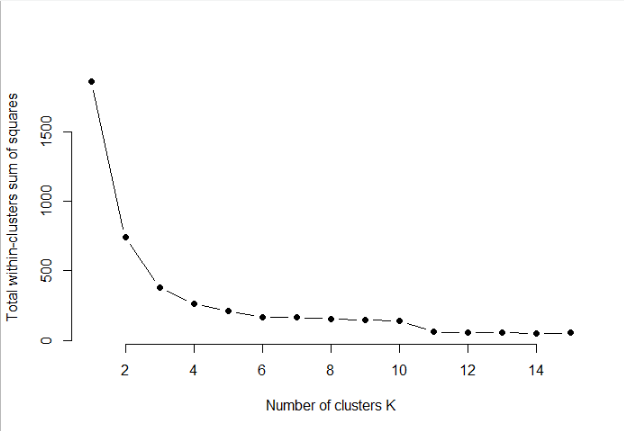


Figure 5 A scree plot to determine the optimal number of k

The three resulting clusters are shown in figure 6. Based on the figure, the first cluster has a mean of approximate 72.65% of the internet user, 18972.12 secure servers and 111.54 subscriptions. The second cluster has a mean of approximately 57.07 % of the internet user, 374368.06 secure severs, and 139.04 subscriptions, while the third first cluster has a mean of approximately 2.31 % of the internet user, 0.00 secure servers and 4.40 subscriptions. The results suggest that the data can be grouped into three distinct groups or digital eras.

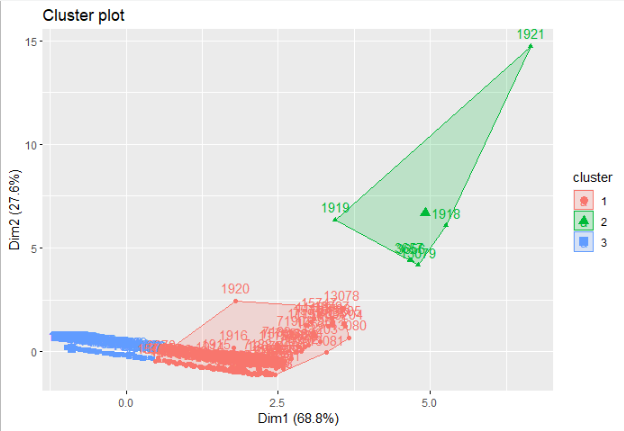


Figure 6 Scatter plot showing the three data clusters

## Linear model

A linear model was conducted to estimate the percentage of the population using the internet based on the mobile cellular subscription and the number of secure internet users. The estimated linear equation is defined as . The results suggest that, on average, 1.91% of the population is expected to use the internet. This is expected to increase by 0.59% if a unit increases the mobile cellular subscription, and a decrease of 0.000049% in the secure server is increased by unit measurement. Overall, the model was found to be significant while it explains 78.6% of the percentage of the internet user variability.

Furthermore, the results in figure 7 show that the linear model assumed linear association, there was no multicollinearity, and the residuals were approximately normal.

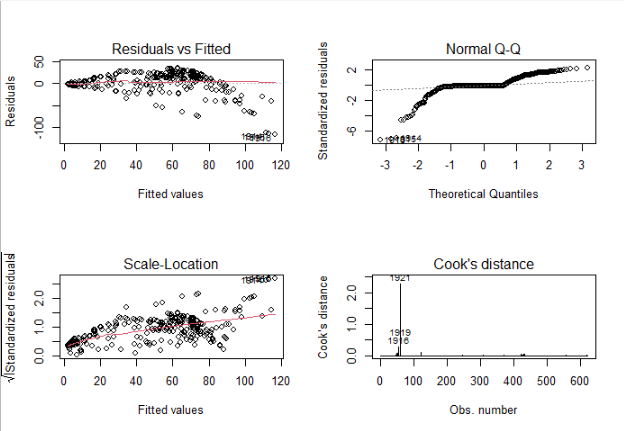


Figure 7 Model diagnostic lots

# Conclusion and Limitations

The current paper seeks to provide insight into the world bank's digital development initiative, which focuses on expanding access to fast, affordable internet. Besides, the study seeks to develop a reliable recommendation for improving service delivery, good governance and social accountability. To achieve this, the study ought to develop an initial study that would determine the impact of the number of secure internet servers and mobile cellular subscriptions on the total percentage of the population that are internet users. Clearly, the results in the descriptive statistics show an upward trend in the three digital parameters. Noticeably, the increase is much noticed as from 1990 to date. This suggests a significant improvement in digital technologies.

The study clearly shows the impact of the technological improvement that has been recently launched. Considering the clustering of the data. Three clear clusters were assumed. This suggests three digital eras. In this context, the study ought to baptize the three eras as; analogue, digital and post-digital eras (Rahmatullah, 2022). During the analogue era, internet connectivity was poor. At this time, the number of secure systems was also small; however, the technology to challenge their security was also poor. During the digital era, there is a need to embrace change, which comes with a technological push (Volberda, 2021). The push led to increased internet users, increased subscriptions, and the need for secure servers. In the postmodern era, data security becomes a priority, and so does the subscription increase. This led to an increase in the number of internet users due to the business push.

In conclusion, the linear model predicted the internet users' percentage significantly. However, the current study was limited to several factors.

1. Small sample size by only considering ten countries.
2. The study was limited to secondary data.

Thus, the study would recommend further studies on the acceptability of the digital presence and the challenges associated with digital developments.

# References

Bayar, Y. G. M. D. &. P. D., 2021. . Impact of mobile phones and internet use on financial inclusion: Empirical evidence from the EU post-communist countries. *Technological and Economic Development of Economy,* pp. 27(3), 722-741.

Bun, M. J. &. H. T. D., 2019. OLS and IV estimation of regression models, including endogenous interaction terms. *Econometric Reviews,* pp. 38(7), 814-827.

David, O. O. &. G. W., 2020. Information and communication technology penetration level as an impetus for economic growth and development in Africa. *Economic research-Ekonomska istraživanja,,* pp. 33(1), 1394-1418..

Embaye, W. T. Z. Y. A. &. C. B., 2021. Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. *Plos one,* pp., 16(2), e0244953.

Flatt, C. &. J. R. L., 2019. Principle assumptions of regression analysis: Testing, techniques, and statistical reporting of imperfect data sets. *Advances in Developing Human Resources,* pp., 21(4), 484-502.

Isaac, O. A. Z. R. T. &. M. A. M., 2018. International Journal of Technological Learning, Innovation and Development, *Factors determining user satisfaction of internet usage among public sector employees in Yemen.,* pp. 10(1), 37-68.

Lee, M. Y. J. J. P. A. W. D. K. F. S. G. .. &. Z. X., 2018. How to respond to the fourth industrial revolution or the second information technology revolution? Dynamic new combinations between technology, market, and society through open innovation. *Journal of Open Innovation: Technology, Market, and Complexity,* pp. 4(3), 21.

Maleki, M. M. M. R. W. D. &. P. K. H., 2020. Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel medicine and infectious disease,* pp. 37, 101742.

Matthess, M. &. K. S., 2020. Structural change and digitalization in developing countries: Conceptually linking the two transformations. *Technology in society,* pp. 63, 101428.

Mattsson, L. G. &. A. P., 2019. Private-public interaction in public service innovation processes-business model challenges for a start-up EdTech firm. *Journal of Business & Industrial Marketing.*

Min, S. S. K. K. F. &. J. M., 2021. Consumer adoption of the Uber mobile application: Insights from the diffusion of innovation theory and technology acceptance model. *Future of Tourism Marketing Routledge.,* pp. (pp. 2-15).

Rahmatullah, A. S. M. E. S. S. P. F. &. P. R. E., 2022. Digital era 4.0: The contribution to education and student psychology. *Linguistics and Culture Review,* pp. 6, 89-107.

Talebian, A. &. M. S., 2018. Predicting the adoption of connected autonomous vehicles: A new approach based on the diffusion of innovations theory. *Transportation Research Part C: Emerging Technologies,* pp. 95, 363-380.

Volberda, H. W. K. S. B.-F. C. M. O. R. &. B. J., 2021. Strategizing in a digital world: Overcoming cognitive barriers, reconfiguring routines and introducing new organizational forms. *Long Range Planning,* pp. 54(5), 102110.

Yang, F. &. Z. X., 2019. . Focal fields in the literature on the information divide: the USA, China, UK and India. *Journal of Documentation.*

# Appendix

# cleaning the memory

rm(list = ls())

#

#setting working directory

setwd("C:/Users/user/Desktop/canny brain/Mee/August/G124691 V1")

# libraries

library(tidyverse)

library(ggplot2)

library(MASS)

library(forecast)

library(reshape2)

library(reshape)

library(lubridate)

# loading the data

users<-read.csv("users.csv")

server<-read.csv("server.csv")

Subscriber<-read.csv("Subscriber.csv")

# melting the data

userM<-melt(users, id = c("Country.Name"))

serverM<-melt(server, id = c("Country.Name"))

SubscriberM<-melt(Subscriber,id = c("Country.Name"))

# renaminin the column names

colnames(userM)<-c("country","year","users")

colnames(serverM)<-c("country","year","servers")

colnames(SubscriberM)<-c("country","year","subscribers")

# merging the two data

data<-merge(x = userM, y = c(serverM,SubscriberM), by = c("country","year"),

all.x = TRUE)

data<-data[,c(1:4,7)]

# removing X in the column year

data$year = as.numeric(gsub("\\X", "", data$year))

# replacing na with 0

data[is.na(data)] <- 0

# selecting the ten countries

data<- data[data$country %in% c("British Virgin Islands","Denmark","Seychelles",

"Netherlands","Singapore", "United States",

"North America","Switzerland","Ireland","Iceland"), ]

# Summary of the data

summary(data)

# top ten countries in the user, subscribers, and servers

df<-data %>% group\_by(country) %>%

summarise\_at(vars(users),list(sum=mean))%>%

arrange(desc(sum))

ggplot(data = head(df,10),aes(x=reorder(country,-sum),y=sum))+

geom\_bar(stat = 'identity')+xlab("Countries ")+

theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))+

ylab(" Total percentage of internet users")

df<-data %>% group\_by(country) %>%

summarise\_at(vars(servers),list(sum=mean))%>%

arrange(desc(sum))

ggplot(data = head(df,10),aes(x=reorder(country,-sum),y=sum))+

geom\_bar(stat = 'identity')+xlab("Countries ")+

theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))+

ylab(" Secure internet servers")

df

df<-data %>% group\_by(country) %>%

summarise\_at(vars(subscribers),list(sum=mean))%>%

arrange(desc(sum))

ggplot(data = head(df,10),aes(x=reorder(country,-sum),y=sum))+

geom\_bar(stat = 'identity')+xlab("Countries ")+

theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))+

ylab(" Mobile Cellular subscribers")

df

# scatterplot

# Subscriber

par(mfrow=c(2,2))

df<-data %>% group\_by(year) %>%

summarise\_at(vars(subscribers),list(sum=mean))%>%

arrange(desc(sum))

plot(df$year,df$sum,xlab = "Year",ylab = "Average mobile cellular Subscribers")

# users

df<-data %>% group\_by(year) %>%

summarise\_at(vars(users),list(sum=mean))%>%

arrange(desc(sum))

plot(df$year,df$sum,xlab = "Year",ylab = "% of internet users")

# Servers

df<-data %>% group\_by(year) %>%

summarise\_at(vars(servers),list(sum=mean))%>%

arrange(desc(sum))

plot(df$year,df$sum,xlab = "Year",ylab = "Secure Internet servers")

dev.off()

# Kmean clustering

# Scalling the data

df<-scale(data[,-c(1:2)])

library(factoextra)

library(cluster)

# determining the optimal number of clusters

set.seed(123)

# function to compute total within-cluster sum of square

wss <- function(k) {

kmeans(df, k, nstart = 10 )$tot.withinss

}

# Compute and plot wss for k = 1 to k = 15

k.values <- 1:15

# extract wss for 2-15 clusters

wss\_values <- map\_dbl(k.values, wss)

plot(k.values, wss\_values,

type="b", pch = 19, frame = FALSE,

xlab="Number of clusters K",

ylab="Total within-clusters sum of squares")

# four

# Kmeans

kmean<- kmeans(df,centers = 3,nstart = 20)

kmean

# plot of the kmean

fviz\_cluster(kmean,data = df)

#find means of each cluster

aggregate(data[,-c(1:2)], by=list(cluster=kmean$cluster), mean)

# predictive model

fit<-lm(users~subscribers+servers,data = data)

#summary

summary(fit)

# model diagonistics

par(mfrow=c(2,2))

plot(fit,which = 1:4)