LINEAR STATISTICAL MODEL PROJECT

TABLE OF CONTENTS

INTRODUCTION

The telecommunications industry in the United Kingdom (UK) is responsible for keeping the UK's 68.9 million inhabitants connected with each other, and the world. Despite some periods of decline, the industry still generated revenue amounting to more than 31 billion British pounds in 2020.[1] The telecommunications industry contributes 4-5% of the UK's GDP and plays a vital role in the economic and social well-being of the country.[2]. Reports also suggests that nine in 10 people think a broadband connection is essential to their everyday life alongside other essentials such as food, housing and utilities like water and energy.[3] The company's market stronghold on broadband and 'Pay as you go' connections would also be beneficial in this situation. Even-though it is considered as a saturated market, we can find possibilities to successfully initiate operations in the U.K and become a strong market competitor if we carry out a meticulous analysis on the user data of the telecommunication sector.

The primary focus of this report is to find out whether it is feasible for the organization to commence its operation in UK on the basis of statistical analysis based on user data of recent mobile users in the country based on different classification of groups. Also, it aims to put forward a recommendation to provide an idea about which group should be given more focus during the formulation of the company's market strategy and on advertisements.

ANALYSIS

PRIMARY DESCRIPTION

The public data provided is divided into 5 data-sets. They are categorized on the basis of Age-group, gender, disability, ethnicity and economic activity. This contains data for people who subscribed to an internet connection for the first three months of years starting from 2014-2020. The main statistical analysis techniques used in this project to produce results are Simple linear regression and its derivatives (Multiple linear regression, Polynomial regression). Number of users for the next 5 years were predicted using the given data. The quality of regression is measured by different statistical estimates associated with linear regression such as residual standard error, adjusted R^2 and P-value of the estimate.

Simple linear regression is a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).[4] A simple linear regression model is represented by the equation "y = B0 + B1*x" where B0, B1 are the coefficients of linear regression. The main objectives of regression are prediction of future observations, assessment of the effect of or relationship between explanatory variables and the response variable and to provide a general description of data structure. The best fit line for the data is found by the sum of squares method where the vertical distance between the observed points and the line is minimum. It is given by the equation sum-squared= $\Sigma i_{=1ton}$ $(y_i - B0 - B1*x_i)^2$where 'y_i' is the predicted variable for the i^{th} observation.

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable.   The formula for a multiple linear regression is: y=B0+B1*XI+….+BnXn where y = the predicted value of the dependent variable, B0 = the y-intercept , B1X1= the regression coefficient (B1) of the first independent variable (X1)[5]. P- value and the residual square error are used to determine the quality of this model.

Polynomial Regression is a form of regression analysis in which the relationship between the independent variables and dependent variables are modelled in the nth degree polynomial.

Polynomial Regression models are usually fit with the method of least squares. The least square method minimizes the variance of the coefficients.[6] The regression equation is given by y= BO+B1*Xi+B2*Xi^2.

Three estimates were used to define the accuracy of the model:

1. Residual error- A residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value.[7]
2. Adjusted R^2- it is a measure obtained from the regression model which explains how much variation in the given data is explained by the given model.
3. P-values and coefficients in regression analysis work together to tell you which relationships in your model are statistically significant and the nature of those relationships. The p-values for the coefficients indicate whether these relationships are statistically significant. If the p-value for a variable is less than your significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population. the data favour the hypothesis that there is a non-zero correlation[8]

ANALYSIS BASED ON AGE GROUP

The recent internet users are categorized based on their age group from 16 to 75 in intervals of 8. The main types of statistical techniques used to find the best prediction of data was linear regression and polynomial regression. The predicted variable(y) is the number of users for the next 5 years from 2021 and the explanatory variable ‘x’ is ‘year’. The table below provides the details about the regression equation used and the corresponding estimate values for each age group.

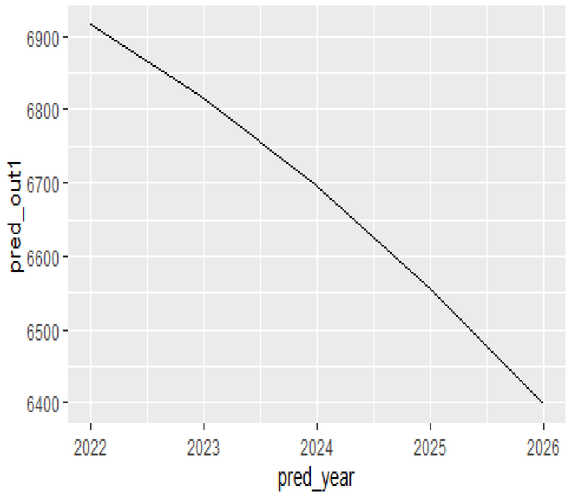
Age group	Residual error	Adjusted R^2	Regression equation	P value
16-24	39.3	0.67	B0+B1*Year+B1*Year^2	0.00157
25-34	86.4	86.4	B0+B1*Year^2	0.000681
35-44	58.27	0.65	B0+B1*Year	0.0199
45-54	18.18	0.98	B0+B1*Year+B1*Year^2	1.46e-06
55-64	19.35	0.98	B0+B1*Year+B1*Year^2	5.907e-07
65-74	34.48	0.998	B0+B1*Year+B1*Year^2	1.707e-07
75+	53.98	0.97	B0+B1*Year	7.88e-06

From this table, we can find that the age groups 45-54, 55-64,65-74, 75+ have the best regression prediction as compared to other age groups. This was concluded on the basis of relatively smaller values of their statistical estimates for accuracy. The graphs of the predicted values based on age groups are given below

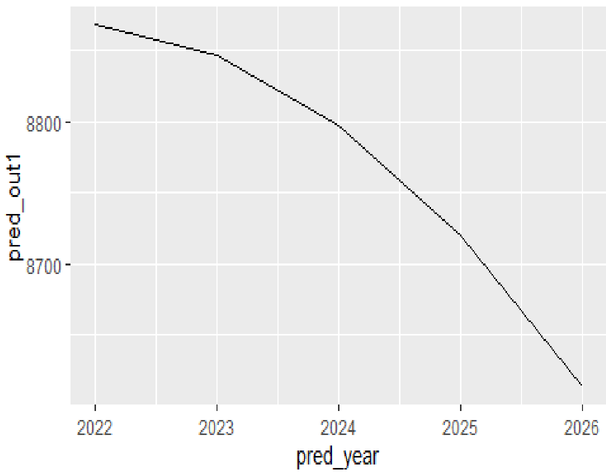
Figures based on age group

X axis= Predicted years

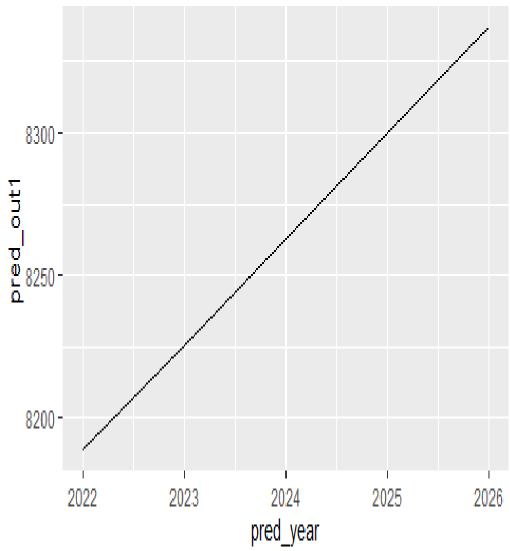
Y axis= Predicted users



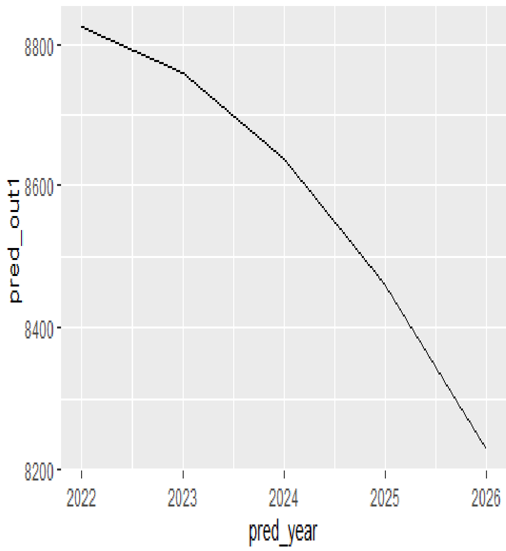
Age group 16-24



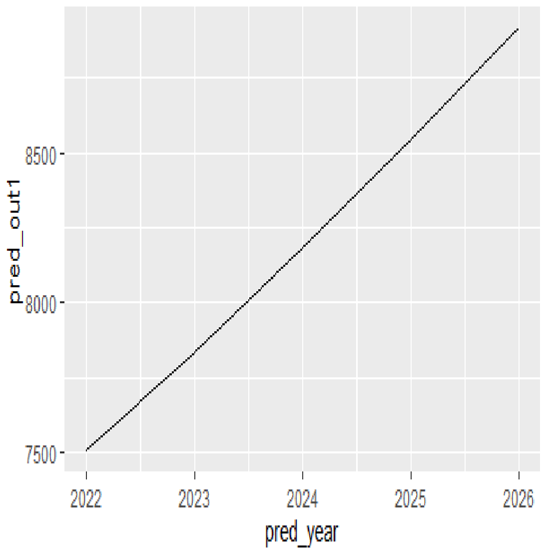
Age group 25-34



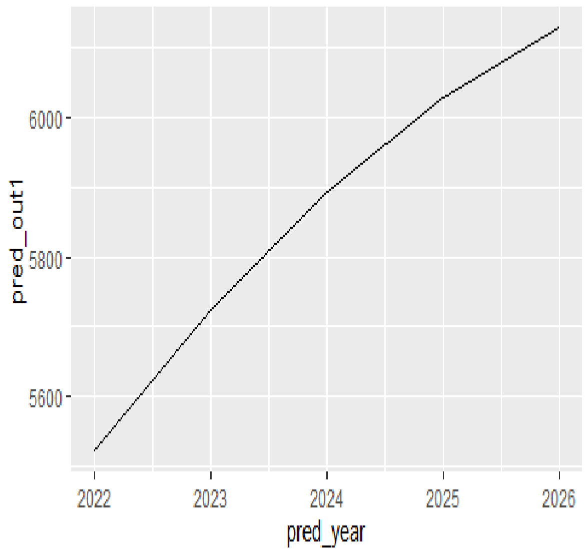
Age group 35-44



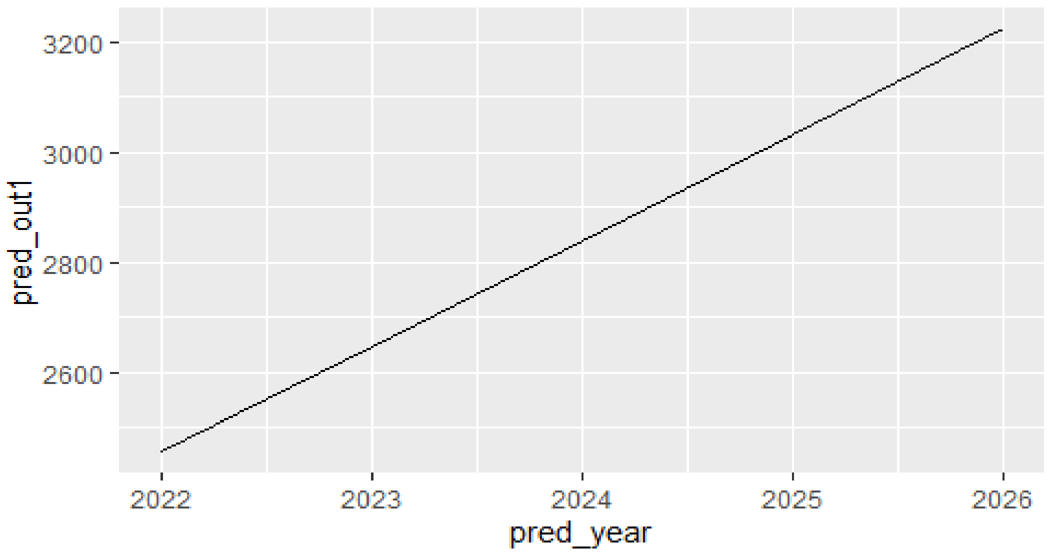
Age group 45-54



Age group 55-64



Age group 65-74



Age group 75+

From the graphs, it is clear that age groups 35-44, 55-64, 65-74, 75+ show an increasing trend in the number of users for the next 5 years. The highest number of new users is predicted in the 55-64 age group. Moreover, these age groups tend to have a positive correlation coefficient(σ) which means that the variables have a positive linear relationship between each other as they tend to increase together. Age groups 16-24,25-34 and 45-54 have a negative correlation coefficient as they tend to decrease together.

ANALYSIS BASED ON AGE GROUP AND GENDER

In this data, the number of new male and female users for each group was provided. Multiple linear regression was used to predict the future number of male and female users. The predictor variables were the number of men or women in a particular age group. The explanatory variables that were taken to account were ‘years’ and ‘total number of users’ (male+female for each age group) which was taken from the previous dataset. The analysis was done separately for Men and women. The following table provides the details on the equation and the estimates for Men:

Age group	Residual error	Adjusted R^2	Regression Equation	P value
16-24	8.294	0.88	BO+B1*Years+B2*(Total number of users)	0.0054
25-34	29.78	0.92	BO+B1*Years+B2*(Total number of users)	0.0025
35-44	7.859	0.9601	BO+B1*Years+B2*(Total number of users)	0.0007055
45-54	12.15	0.9962	BO+B1*Years+B2*(Total number of users)	6.32e-06
55-64	11.05	0.9982	BO+B1*Years+B2*(Total number of users)	1.454e-06
65-74	10.2	0.98	BO+B1*Years+B2*(Total number of users)	3.671e-07
75+	13.48	0.99	BO+B1*Years+B2*(Total number of users)	1.155e-05

From the table, we can see that the model could predict the data for each group by a good accuracy. The variance in the data could be predicted by the model and the variables used are statistically significant.

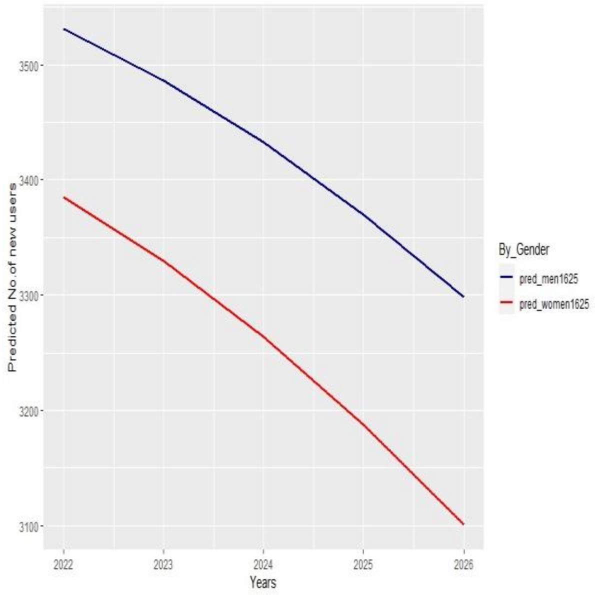
The following table provides the details on the equation and the estimates for Women:

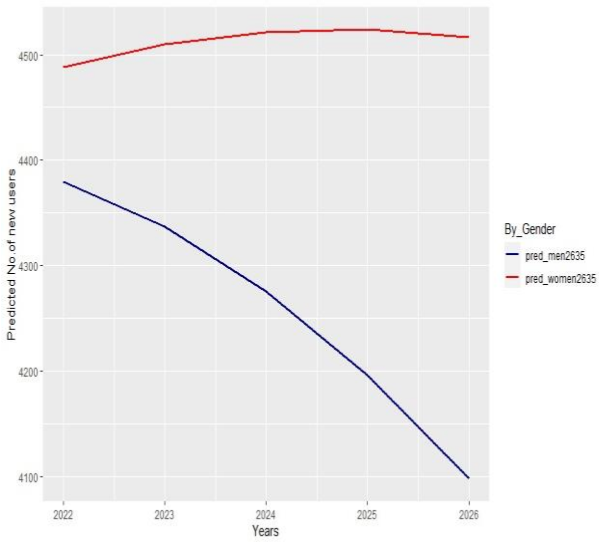
Age group	Residual error	Adjusted R^2	Regression equation	P value

16-24	8.233	0.9304	BO+B1*Years+B2*(Total number of users)	0.002151
25-34	29.78	0.9584	BO+B1*Years+B2*(Total number of users)	0.0007695
35-44	7.859	0.9814	BO+B1*Years+B2*(Total number of users)	0.0001536
45-54	12.15	0.99	BO+B1*Years+B2*(Total number of users)	3.498e-06
55-64	11.05	0.9987	BO+B1*Years+B2*(Total number of users)	7.151e-07
65-74	10.2	0.995	BO+B1*Years+B2*(Total number of users)	1.271e-07
75+	13.48	0.98	BO+B1*Years+B2*(Total number of users)	5.323e-08

Similar to the estimates of regression for men, the model provides a good fit for the predicted number of users for women. Compared to the prediction estimates of other age groups, 65-74, 75+ age groups have a relatively better prediction.

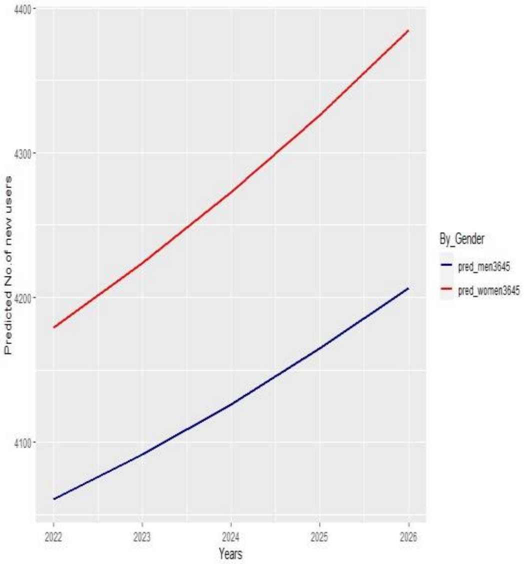
The graphs for the predicted number of users for men and women are given below:



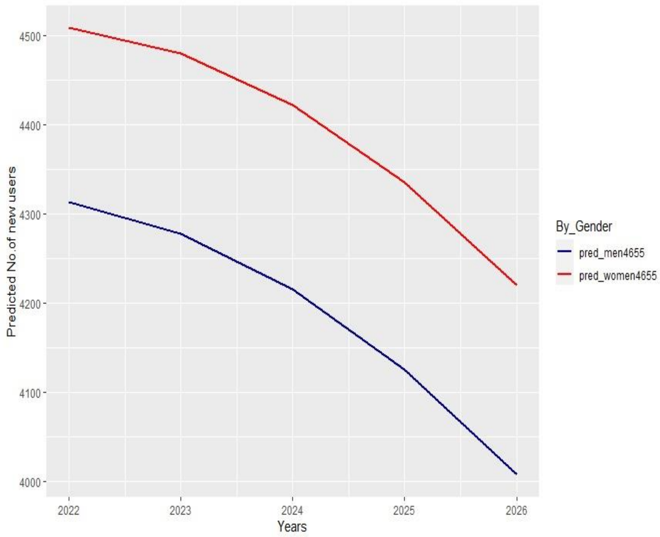


Age group 16-24

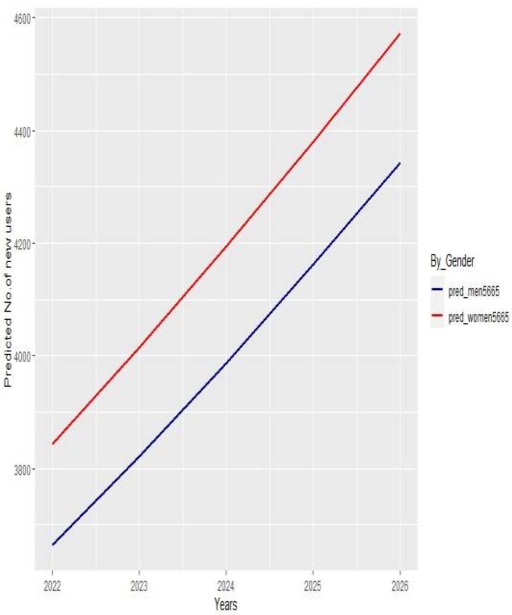
Age group 25-34



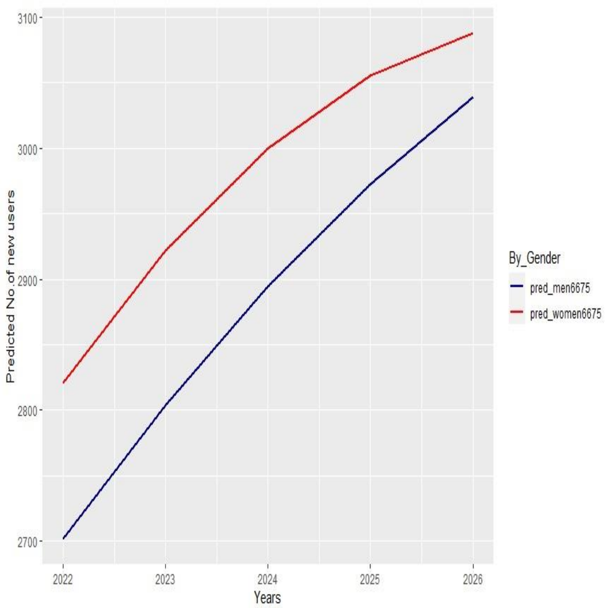
Age group 35-44



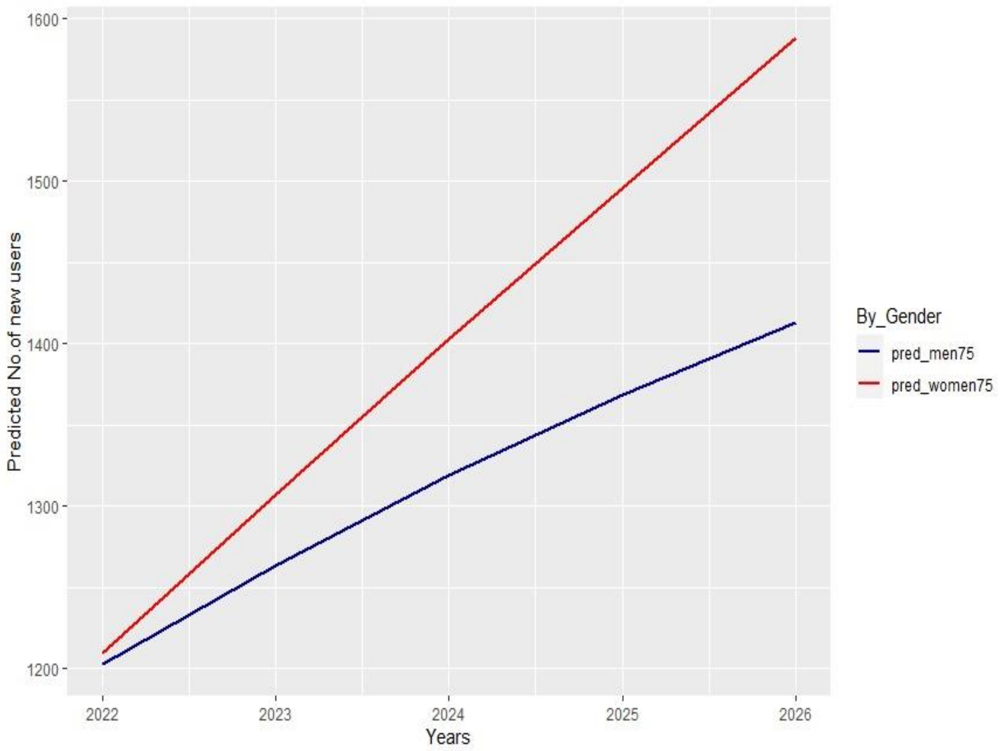
Age group 45-54



Age group 55-64



Age group 65-74



Age group 75+

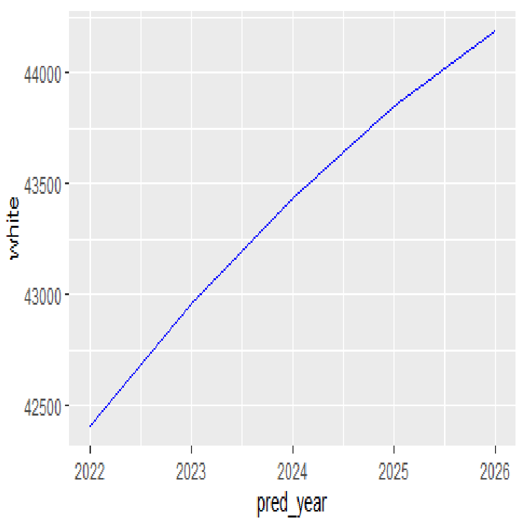
From the graphs we can see that the model predicted an increased number of male and female mobile users for age groups 35-44, 55-64,65-75, 75+. Both predicted number men and women users follow roughly the same kind of linear relationship with to the dependent variables. An interesting trend to be noticed is for the age-group 25-34 as the number of women show a slight increase with time, however the predicted number of men users is set for a steep decrease over the course of next 5 years. The model predicts a steady decrease for Women and men users in group 45-54. In general, we can see that at any point in time for the predicted users, number of women are significantly higher than the potential number of new men users.

ANALYSIS BASED ON ETHNICITY

The data contains information about the new internet users for the first three months of years starting from 2014-2021 based on different ethnic groups in the UK. Polynomial regression was used to fit the data to predict the values of new users across different ethnicities. The explanatory variable taken was ‘year’ and the predicted variable was the number of new users for each ethnic group.

Ethnicity	Residual error	Adjusted R^2	Regression equation	P value
white	149.9	0.99	B0+B1*Year+B2*Year^2	5.987e-07
Mixed/multiple ethnic background	35.49	0.7843	B0+B1*Year+B2*Year^2	0.00931
Indian	20.82	0.9471	B0+B1*Year+B2*Year^2	0.0002773
Pakistani	7.618	0.9751	B0+B1*Year+B2*Year^2	7.25e-07
Bangladeshi	15.31	0.9169	B0+B1*Year+B2*Year^2	0.0008577
Other Asian background	28.5	0.7732	B0+B1*Year+B2*Year^2	0.01056
Black/African/Caribbean/Black British	49.07	0.9134	B0+B1*Year+B2*Year^2	0.000952
Chinese	12.35	0.7831	B0+B1*Year+B2*Year^2	0.009123
Other ethnic group	36.63	0.7887	B0+B1*Year+B2*Year^2	0.008951

From the table, we can infer that the model produces reasonable fit for the given data. We can see that the model fits the best for data for Pakistan as its has the best value for estimates of regression. The corresponding graphs for the predicted values are given below:



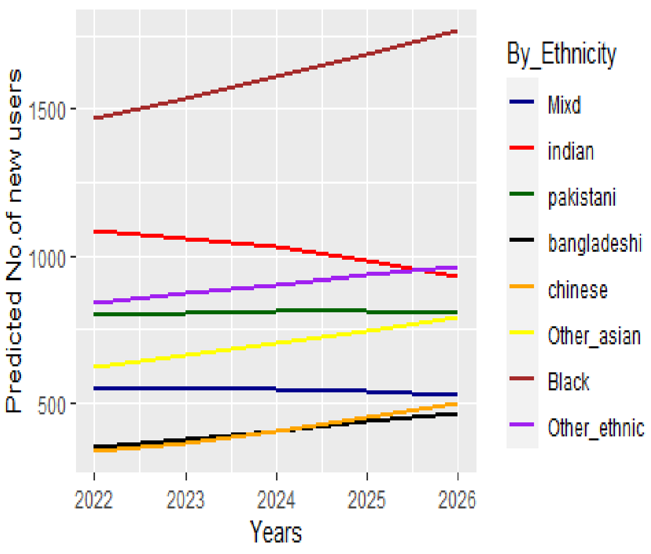


Fig: Predicted white users over next 5 years

Fig: Predicted new users of different ethnicity

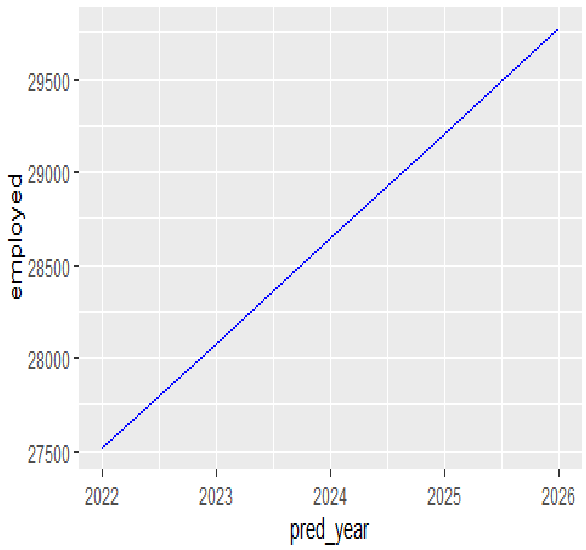
As expected, the number of white users is set to increase linearly for the given time period. It has a positive correlation coefficient. Among other ethnicities, The Black/African/Caribbean/Black British population dominates the trend by a significant difference Whereas other ethnicities show a relatively small growth over time except the Indian and mixed ethnicities. They show a decreasing trend over the course of next 5 years. The cumulative number of ‘other ethnic groups’ will overcome the number of Indian users. Bangladeshis would have the least number of new users in the next 5 years.

ANALYSIS BASED ON ECONOMIC ACTIVITY

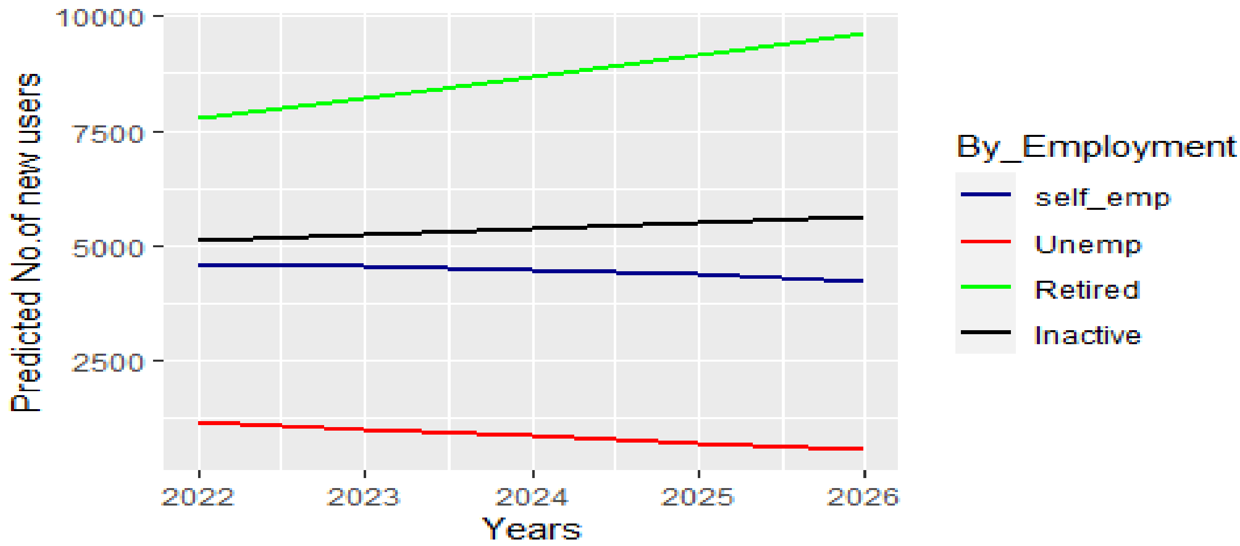
Simple linear regression and polynomial regression were used to model the given data. The regression equation was formed with the ‘year’ as the explanatory variable and ‘number of new users by each economic stream’ as the predictor variable and the analysis was done. The following table provides details on the regression analysis done.

Economic activity	Residual error	Adjusted R^2	Regression equation	P value
Employee	154.3	0.9876	$B0+B2*year^2$	3.754e-07
Self-employed	84.75	0.9573	$B0+B2*Year^2+B3*Year^3$	0.0001624
Unemployed	98.45	0.88	$B0+B2*year^2$	0.0003244
Retired	37.96	0.9989	B0+B1*Year	2.84e-10
Inactive	82.72	0.9392	$B0+B2*Year^2$	4.509e-05
Unpaid family worker	8.675	0.4406	B0+B1*Year	0.0315

The following are the graphs of the analysis:



Predicted graph for employee population   Predicted graph for unpaid workers



From the graphs, we can see that majority of the economic activity groups produce an increasing trend in number of new predicted users except Unemployed and inactive users which is reasonable. The employed population is set to have a steep increase in the number of new predicted users and has positive correlation coefficient. This group would have the highest number of new users over the course of five years. Surprisingly, the retired population would have the second highest predicted new internet users. Another surprising trend is shown by the unpaid family workers which predicts a significant growth in the next 5 years. There is a linear relationship between the variables for this group. As expected, the unemployed population would have considerably less number of new users than other economic groups.

The analysis for the student population and the government workers were not satisfactory using the recommended regression models as they produced unsatisfactory estimates of regression thus leading to bad predictions. A more reasonable model could be produced if more data is available on the discusses population for a more fruitful analysis.

RECOMMENDATION

From the analysis of the given data and based on the results produced the following recommendation is made to the board. The organization's market expansion to the United Kingdom would be only profitable and reasonable if the company focus on certain categories of population rather than attending to the whole population. The main type of connections company should focus on should be broadband connections. The reasons for that are explained below

From the analysis of age group, it was found that the number of new internet users in the senior population especially for the age groups ranging from 55-75 is set to have a significant increase. As the senior population is more likely to subscribe to a broadband connection rather than a pay as you go connection or a prepaid connection, more attractive plans and offers for broadband connections should be marketed. It would be fair to think less focus might be given to the younger population as other companies provide better plans and offers for them. The money and time invested to market plans among them should be minimal.

It was found that ,although women and men followed the dame trend majority of the age groups, the difference between new women and men internet users is significant. The reason behind this is unknown as more research is required to study this observed dynamic for a concrete explanation. Introducing plans and subscription offers favouring women might give the company a slight edge of profit as observed from the given data. The age- group that should be more focussed on this recommendation would be the 25-34 group as it depicts the highest difference between new women and men users.

It was found that the equality act internet users' data was insignificant for the prediction. It would not be profitable for the company to invest resources to this population as they receive highly benefitted offers from the government. Moreover, the data given does not provide a distinct picture as a number of respondents who chose not to declare whether they had a disability have been included within the category 'Not Equality Act’. A proper prediction could not be made without more accurate data.

Analysing users based on ethnicity gave a clear message on inclusivity. it was found that almost all the ethnic groups are predicted to have an increased user base of internet. This gives us the message that advertisement should be more targeted on these groups. More inclusive and personalized adverts for different ethnic groups would be way to increase the company’s popularity among these ethnic groups. Eventhough the number of white people have a vast difference in user basis, the cumulative sum of all the new users in ethnic groups counterbalance this difference.

The analysis on new users based on the economic activity gives us new evidence on the idea why the company should focus on broadband connections. From the data it was seen that, employee population and the retired population topped the charts for the analysis. These are two groups who heavily rely on broadband connections in their homes or offices. Another recommendation to be made is the introduction of more interactive post-paid plans with provisions like unlimited data or calls as this would be greatly beneficial for the employee population.

These are the few recommendations based on the insights from the analysis of data. The accuracy of the analysis could be greatly improved if more information is provided on the lapsed internet users, thus leading to a more precise and targeted market strategy for the expansion.

REFERENCES

1- https://www.statista.com/topics/7198/telecommunications-industry-in-the-united-kingdom/#dossierKeyfigures

2 https://www.moorhouseconsulting.com/insights/perspectives/current-uk-network-landscape-a-flat-market-and-tighter-margins/

3- https://press.which.co.uk/whichpressreleases/broadband-now-seen-as-one-of-top-five-modern-day-essentials/

4-https://machinelearningmastery.com/linear-regression-for-machine-learning/

5-https://www.scribbr.com/statistics/multiple-linear-regression/

6- https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18

7-https://towardsdatascience.com/how-to-use-residual-plots-for-regression-model-validation-c3c70e8ab378

8-https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

APPENDIX

```
library(tidyverse)

library(eeptools)

library(ggplot2)

library(readxl)



df<- read_excel('LSM.xlsx')


colnames(df) <- c("V2","V3","V4","V5","V6","V7","V8","Year")

df <- as.data.frame(t(df))

df <- df[-c(1), ]


df$Year<- c(2015,2016,2017,2018,2019,2020,2021)

model<- lm(V7~Year, data=df)

summary(model)

a<-coef(model)["(Intercept)"]

b<-coef(model)["Year"]

b2<-coef(model)["I(Year^2)"]


pred_year<- c(2022,2023,2024,2025,2026)

pred_out<-numeric(5)

for (x in pred_year){

  pred_out<- c(pred_out,a+(b*x))

}

print(pred_out)

pred_out1<-c( 2456.429,2648.000, 2839.571 ,3031.143 ,3222.714)


plot_data<- data.frame(pred_year,pred_out1)
```

```
ggplot(plot_data,mapping = aes(x = pred_year , y = pred_out1)) +
  geom_line()


#Prediction based on gender


df_1<- read_excel("LSM_Gen.xlsx")
df_1<- na.omit(df_1)
colnames(df_1) <-
c("V1","V1.1","V2","V2.1","V3","V3.1","V4","V4.1","V5","V5.1","V6","V6.1","V7","V7.1")
df_1 <- as.data.frame(t(df_1))


df_1$Year<- c(2013,2014,2015,2016,2017,2018,2019,2020,2021)
df_1 <- df_1[-c(1,2), ]
rownames(df_1)<- c("1","2","3","4","5","6","7")


model<- lm(df$V1~Year+I(Year^2), data=df_1,)
summary(model)
a<-coef(model)["(Intercept)"]
b<-coef(model)["Year"]
b2<-coef(model)["I(Year^2)"]


#Prediction based on employment


df_2<- read_excel("LSM_emp.xlsx")
df_2 <- df_2[-c(1),]
df_2 <- df_2[-c(2)]
df_2 <- as.data.frame(t(df_2))
colnames(df_2) <- c("Category","V1","V2","V3","V4","V5","V6","V8","V9")
df_2$Year<- c(2014,2015,2016,2017,2018,2019,2020,2021)
df_2 <- df_2[-c(1),]
```

```
model_3<- lm(V4~I(Year^3), data=df_2)
summary(model_3)

a<-coef(model_3)["(Intercept)"]
b<-coef(model_3)["Year"]
b2<-coef(model_3)["I(Year^2)"]



pred_year<- c(2022,2023,2024,2025,2026)
pred_out<-numeric(5)
for (x in pred_year){
  pred_out<- c(pred_out,a+(b2*(x^2)))
}
print(pred_out)



employed<-c(27513.29,28077.73,28642.45,29207.44,29772.72)
self_emp<-c(4604.337 ,4571.756 ,4500.457,4390.400,4241.549)
Unemp<-c(1189.9792,1034.5240,878.9919,723.3829, 567.6970)
Retired<-c(7765.286,8225.905,8686.524,9147.143,9607.762)
Inactive<-c(5117.607,5250.964,5384.321,5517.679,5651.036)
Unpaid<- c( 111.0107,114.4354,117.8618,121.2899,124.7196)

plot_data<- data.frame(pred_year,self_emp,Unemp,Retired,Inactive)
plot_data1<-data.frame(pred_year,Unpaid)
plot_data1_1<-data.frame(pred_year,employed)
ggplot() +
  geom_line(data = plot_data1_1, aes(x = pred_year, y = employed), color = "blue")



#colors<-c("self_emp"="blue","Unemp"="red","Retired"="green","Inactive"="black")
```

```
ggplot() +
  geom_line(data = plot_data, aes(x = pred_year, y = self_emp, color = "self_emp"),size=1) +
  geom_line(data = plot_data, aes(x =pred_year , y = Unemp, color = "Unemp"), size=1)+
  geom_line(data = plot_data, aes(x =pred_year , y = Retired, color = "Retired"),size=1)+
  geom_line(data = plot_data, aes(x =pred_year , y = Inactive, color = "Inactive"), size=1)+
  scale_color_manual(name = "By_Employment", values = c("self_emp" = "purple", "Unemp" =
"blue","Retired"="red","Inactive"="orange"))+
    xlab('Years') +
    ylab('Predicted No.of new users')



#By_ethnicity


df_3<- read_excel("LSM.Ethnic.xlsx")
df_3 <- df_3[-c(2)]
df_3 <- as.data.frame(t(df_3))
df_3 <- df_3[-c(1),]
df_3$Year<- c(2014,2015,2016,2017,2018,2019,2020,2021)

model_4<- lm(V9~Year+I(Year^2), data=df_3)
summary(model_4)

a<-coef(model_4)["(Intercept)"]
b<-coef(model_4)["Year"]
b2<-coef(model_4)["I(Year^2)"]

pred_year<- c(2022,2023,2024,2025,2026)
pred_out<-numeric(5)
for (x in pred_year){
  pred_out<- c(pred_out,a+(b*x)+(b2*(x^2)))
}
```

```
print(pred_out)

white<-c(42402.43,42954.43,43436.00,43847.14,44187.86)

plot_data2<-data.frame(pred_year,white)

ggplot() +

  geom_line(data = plot_data2, aes(x = pred_year, y = white), color = "blue")


Mixd<-c(548.5536,550.8869,548.0417,540.0179,526.8155)

indian<-c(1084.5179,1062.2321,1028.6012,983.6250,927.3036)

pakistani<-c(796.8750,808.3036,813.2440,811.6964,803.6607)

bangladeshi<-c(348.6607,376.3750,405.4345,435.8393,467.5893)

chinese<- c(331.9464,365.8036,404.6964,448.6250,497.5893 )

Other_asian<- c(622.2143,660.0714,701.0714,745.2143,792.5000)

Black<-c(1465.804,1538.923,1613.482,1689.482,1766.923)

other_ethnic<-c(841.5179,871.8274,902.2202,932.6964,963.2560)


plot_data_eth<- data.frame(pred_year,Mixd,indian,
pakistani,bangladeshi,chinese,Other_asian,Black,other_ethnic)



ggplot() +

  geom_line(data = plot_data_eth, aes(x = pred_year, y = Mixd, color = "Mixd"),size=1) +

  geom_line(data = plot_data_eth, aes(x =pred_year , y = indian, color = "indian"), size=1)+

  geom_line(data = plot_data_eth, aes(x =pred_year , y = pakistani, color = "pakistani"),size=1)+

  geom_line(data = plot_data_eth, aes(x =pred_year , y = bangladeshi, color = "bangladeshi"),
size=1)+

  geom_line(data = plot_data_eth, aes(x =pred_year , y = chinese, color = "chinese"), size=1)+

  geom_line(data = plot_data_eth, aes(x =pred_year , y = Other_asian, color = "Other_asian"),
size=1)+

  geom_line(data = plot_data_eth, aes(x =pred_year , y = Black, color = "Black"), size=1)+

  geom_line(data = plot_data_eth, aes(x =pred_year , y = other_ethnic, color = "Other_ethnic"),
size=1)+
```

```
scale_color_manual(name = "By_Ethnicity", values = c("Mixd" = "pink", "indian" =
"green","pakistani"="brown","bangladeshi"="orange","chinese"="black","Other_asian"="brown","Bl
ack"="darkblue","Other_ethnic"="black"))+
xlab('Years') +
ylab('Predicted No.of new users')
```