STATISTICAL ANALYSIS PROJECT

Fariya Bano

University of Leicester

This report contains the analysis and evaluation of the statistical trends of a Telecommunication company that is leading the market in Asia but wishes to expand to the UK.

Table of Contents

Introduction

Statistical Analysis is a key technique to investigate quantitative data in order to gain valuable insights or predictions and uncover patterns and trends. In this particular case study, the data collected represents Internet use in the UK with annual estimates by age, sex, disability, ethnicity and economic activity. The client for this project is a Telecommunications company that is a market leader in Asia, but is now focused on expanding operations globally, with UK being the first destination. The objective is to use the findings to evaluate whether it is feasible to initiate operations and which groups should be the focus of investment and advertising. The questions to be answered through this project are:

1. Is this a profitable objective?
2. Which groups should the marketing be focused towards?

The method of focusing advertising and marketing towards specific target groups is known as Targeted Marketing. It allows the targeted audience to familiarize themselves with the product through specific traits and behaviours. As a Project Executive, demographic are they key factor that decide which groups the product should be advertised towards.

Important Concepts

The results of statistical analysis on the data are obtained using statistical concepts Hypothesis testing, T-test, Analysis of Variance and Linear Regression in R Programming Language. During the initial stage of the project, the most import step is to draw valid conclusions is the planning phase. A correct plan involves specifying hypothesis, testing the hypothesis, finding relationship between variables and choosing inferences that serve the objective.

Hypothesis Testing.

A hypothesis is "a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.". In other words, a hypothesis is an assumption of an experiment result that is tested to see if is true or not. There are two types of hypotheses, null and alternative hypothesis. A null hypothesis states that there is no relationship between variables and is denoted by H_0. While, the alternative hypothesis states the opposite, i.e., that there is some relationship between variables and is denoted by H_1. After stating the null and alternative hypothesis for the experiment, we assume that H_0 is true and then apply a suitable test to find evidence to support (or against) this assumption.

Significance Level.

Significance level is the parameter used to test the probability with which the null hypothesis must be rejected when it is true. It is denoted by $\propto$. In simple terms, it is the probability of saying that a condition or assumption is true when it isn't. Considering $\alpha = 0.05$ means that there exists a risk of 5% of the assumption being true when it is not.

T - Test.

"A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features." A t-test allows us to compare the variables and gives us the probability of the null hypothesis being correct. A T-test is most suitable when the sample data contains a categorical variable of 1 or 2 levels only.

One Way ANOVA

ANOVA is a statistical test which means "Analysis of Variance". It is a measure in the difference of means of variables. One Way ANOVA is specifically used to measure statistical

differences when categorical independent variables of 3 or more levels are involved against a quantitative dependent variable.

Simple Linear Regression

"Simple linear regression provides a model of the relationship between the magnitude of one variable and that of a second." Simple linear regression is modelled as $Y_i = a + bX_i$, where Y_i is the i^{th} response variable. X_i is the i^{th} predictor variable, b is the measure of slope i.e., the change in Y with respect to every 1-unit change in X. The Y-intercept of the line is represented by a. A linear regression line is basically a line that passes through almost every point (X_i, Y_i) on a graph. Being a line, it possesses a slope and Y-intercept. Hence, the simple linear regression equation given earlier where the slope b can also be referred to as the regression coefficient. To test if our linear regression model is creating the best fit line for the data, we use R^2 as the measure.

$$R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ variance}$$

The better the R^2value, the better is the fit of the model on the data. If R^2= 1, then it is the best fit model.

The Data

Understanding the Data

The data used in this project is collected as published by the Office for National Statistics (UK) and provides information about recent and lapsed internet users in the UK. The sampling of the data is done over a period of 8 years, from 2014 to 2021, considering only the first 3 months of each year. Recent internet users are adults who have used the internet within the last 3 months. Lapsed internet users are adults who used the internet more than 3 months ago. Information in the published data is broadly divided into 6 datasets and then used for analysis. The Datasets are as follows:

1. Recent and lapsed Internet Users by Age-Group: This data measures the number of internet users from 2014- 2021 in different age groups ranging from 16 to 75+. The different age-groups are 16-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75+. *(This dataset will here forth be referred to as Dataset 1)*

	Age	2014	2015	2016	2017	2018	2019	2020	2021
1	16-24	7145	7088	7075	7074	7155	7129	7036	6992
2	25-34	7903	8162	8457	8660	8582	8720	8815	8894
3	35-44	8074	7986	7952	7900	8053	8129	8118	8145
4	45-54	7430	7694	8005	8290	8498	8686	8803	8814
5	55-64	5434	5624	5821	6060	6361	6607	6888	7189
6	65-74	2799	3153	3562	3939	4390	4721	5031	5264
7	75+	898	1057	1371	1534	1632	1925	2050	2262

Figure 2.1. Recent and lapsed Internet Users by Age – Group data.

2. Recent and lapsed Internet Users by their Sex: This data measures the number of Males and Females that availed Internet Usage from 2014 to 2021. *(This dataset will here forth be referred to as Dataset 2).*

	Sex	2014	2015	2016	2017	2018	2019	2020	2021
1	Men	20039	20564	21242	21814	22229	22812	23200	23606
2	Women	19645	20200	21001	21643	22442	23105	23542	23954

Figure 2.2. Recent and lapsed Internet Users by Age – Group and Sex data

3. Recent and lapsed Internet Users by Sex in each Age- Group: The number of internet users is divided into Age Groups, within which they are further subcategorized as Male or Female. *(This dataset will here forth be referred to as Dataset 3).*

	AgeGroup	Sex	2014	2015	2016	2017	2018	2019	2020	2021
1	All	Men	20039	20564	21242	21814	22229	22812	23200	23606
2	All	Women	19645	20200	21001	21643	22442	23105	23542	23954
3	16-24	Men	3643	3610	3593	3590	3638	3622	3594	3561
4	16-24	Women	3503	3477	3482	3484	3517	3507	3443	3431
5	25-34	Men	3989	4136	4272	4408	4276	4340	4402	4454
6	25-34	Women	3914	4026	4186	4252	4307	4380	4413	4440
7	35-44	Men	4010	3969	3939	3925	3975	4018	4011	4025
8	35-44	Women	4064	4016	4013	3976	4078	4111	4107	4120
9	45-54	Men	3661	3803	3945	4058	4182	4270	4314	4310
10	45-54	Women	3770	3891	4059	4232	4315	4416	4489	4504
11	55-64	Men	2710	2807	2877	3004	3118	3240	3375	3516
12	55-64	Women	2724	2817	2944	3056	3244	3367	3513	3673
13	65-74	Men	1472	1652	1847	1984	2183	2323	2471	2580
14	65-74	Women	1327	1501	1715	1955	2207	2398	2560	2684
15	75+	Men	554	587	769	846	858	998	1033	1159
16	75+	Women	344	471	603	688	774	926	1017	1103

Figure 2.3. Recent and lapsed Internet Users by Sex data

4. Recent and lapsed Internet Users by Disability in each Age-Group: The number of users in each age group are further divided into “Equality Act Disabled” and “Not Equality Act Disabled”. “Equality Act Disabled” refers to those who self-assess that they have a disability in line with the Equality Act definition of disability. A number of respondents who chose not to declare whether they had a disability have been included within the category “Not Equality Act Disabled”. *(This dataset will hereforth be referred to as Dataset 4).*

	AgeGroup	Disability	2017	2018	2019	2020	2021
1	16-24	Equality Act Disabled	697	752	824	834	824
2	16-24	Not Equality Act Disabled	6377	6403	6306	6202	6168
3	25-34	Equality Act Disabled	864	923	1026	1031	1144
4	25-34	Not Equality Act Disabled	7796	7660	7694	7785	7750
5	35-44	Equality Act Disabled	1091	1127	1185	1194	1207
6	35-44	Not Equality Act Disabled	6809	6926	6945	6924	6937
7	45-54	Equality Act Disabled	1421	1562	1591	1634	1732
8	45-54	Not Equality Act Disabled	6869	6936	7095	7169	7082
9	55-64	Equality Act Disabled	1472	1594	1651	1725	1845
10	55-64	Not Equality Act Disabled	4588	4767	4956	5163	5344
11	65-74	Equality Act Disabled	1198	1330	1477	1643	1721
12	65-74	Not Equality Act Disabled	2741	3060	3245	3388	3543
13	75+	Equality Act Disabled	672	751	834	923	1070
14	75+	Not Equality Act Disabled	862	881	1091	1128	1192

Figure 2.4. Recent and lapsed Internet Users by Disability and Age – Group data

5. Recent and lapsed Internet Users by their Ethnicity: The Internet users in UK are broadly categorized on the basis of multiple ethnicities and the corresponding number of users are logged. *(This dataset will hereforth be referred to as Dataset 5).*

	Ethnicity	2014	2015	2016	2017	2018	2019	2020	2021
1	White	35546	36430	37585	38601	39498	40526	40885	41825
2	Mixed/multiple ethnic background	343	401	406	471	471	490	591	510
3	Indian	875	906	954	1025	1086	1090	1094	1087
4	Pakistani	475	527	593	653	684	715	759	779
5	Bangladeshi	171	195	213	239	252	247	316	320
6	Chinese	249	220	250	226	251	267	275	303
7	Other Asian background	423	466	465	467	464	569	562	576
8	Black/African/Caribbean/Black British	940	981	1098	1063	1186	1253	1394	1352
9	Other ethnic group	616	617	655	684	746	711	838	784

Figure 2.5. Recent and lapsed Internet Users by Ethnicity

6. Recent and lapsed Internet Users by their Economic Activity: Various economic activities are used to categorize the Internet Users and their numbers are recorded. *(This dataset will hereforth be referred to as Dataset 6).*

	Activity	2014	2015	2016	2017	2018	2019	2020	2021
1	Employee	23147	23390	24056	24599	25463	25971	26346	26848
2	Self-employed	3493	3788	3865	4266	4288	4510	4621	4543
3	Government employment & training programmes	118	87	144	116	94	86	92	57
4	Unpaid family worker	84	82	84	107	107	90	105	106
5	Unemployed	2217	2363	2318	2061	1719	1610	1479	1340
6	Student	2470	2426	2422	2428	2460	2391	2443	2459
7	Retired	4084	4493	5047	5497	5885	6404	6812	7318
8	Inactive	4070	4134	4307	4383	4658	4855	4844	4889

Figure 2.6. Recent and lapsed Internet Users by Economic Activity

Summarizing the Data

After careful inspection of data, the data can be summarized and visualized using measures of central tendency: mean, median and mode.

1. Mean: The average of all values calculated by the summation of all values divided by the number of values in a given dataset.
2. Median: The value in the exact middle when all values are arranged in ascending order.
3. Mode: The value which is mostly commonly occurring in the entire dataset.

Dataset 1: Age Data.

Figure 2.7 shows a bar graph for data of each of the age-groups in the given dataset. From this graph, notice the variation in the number of users in each group. Also, notice that we can make an inference from the graph that some age groups use internet more than others.

Figure 2.7. Bar Plot of Age Data for each year

To view this difference between Age-Group a Pie chart gives a clearer understanding by taking mean of the data across the years. It appears clear the age groups 65-74 and 75+ have much less internet usage figures hen compared with the other age-groups.

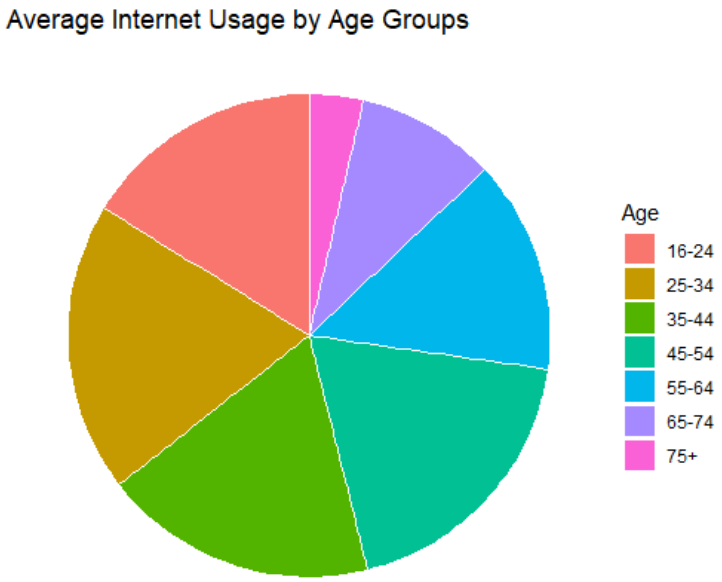

Figure 2.8. Pie Graph of mean data of each Age – group.

A better understanding of the variations in the data can be gained from a boxplot. We see that the previously stated inference appears to be true. We can also notice the top 3 age – groups with most number of internet users, 25 -34, 35-44, and 45-54.

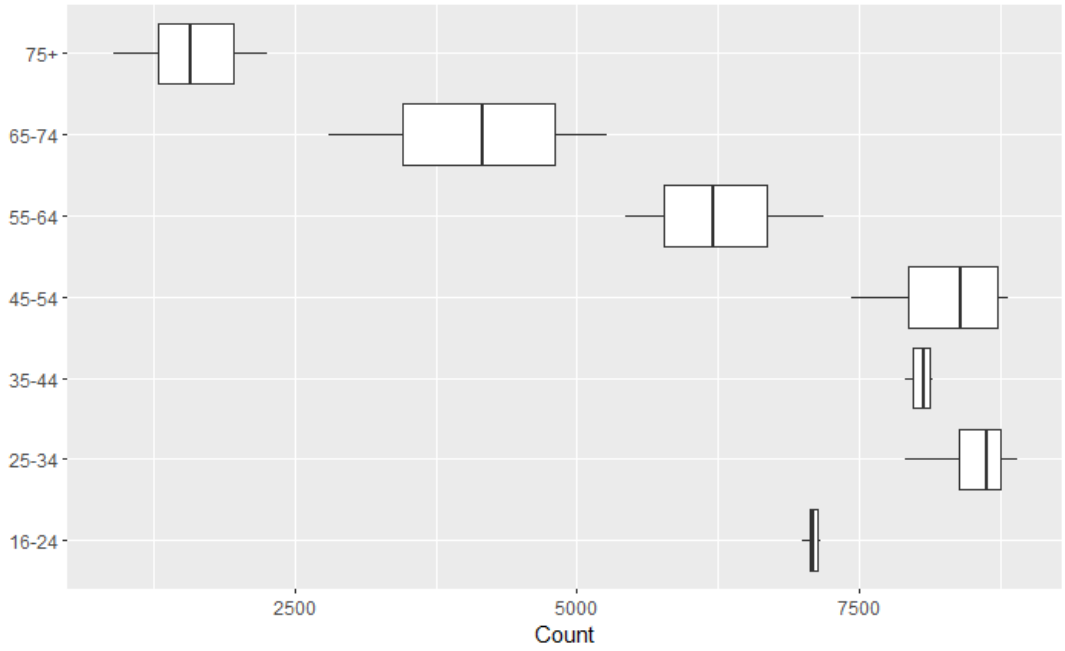

Figure 2.9. Box plot of Internet Users by Age from 2014 – 2021

Dataset 2: Sex Data.

Figure 2.10 shows bar graphs when data is plotted between the 2 levels of Sex variable, Men and Women. When plotted together and when plotted separately, there is a steady rise in the number of users for both. When plotted together, it appears like the number of men using internet is quite similar to the number of women using the internet.

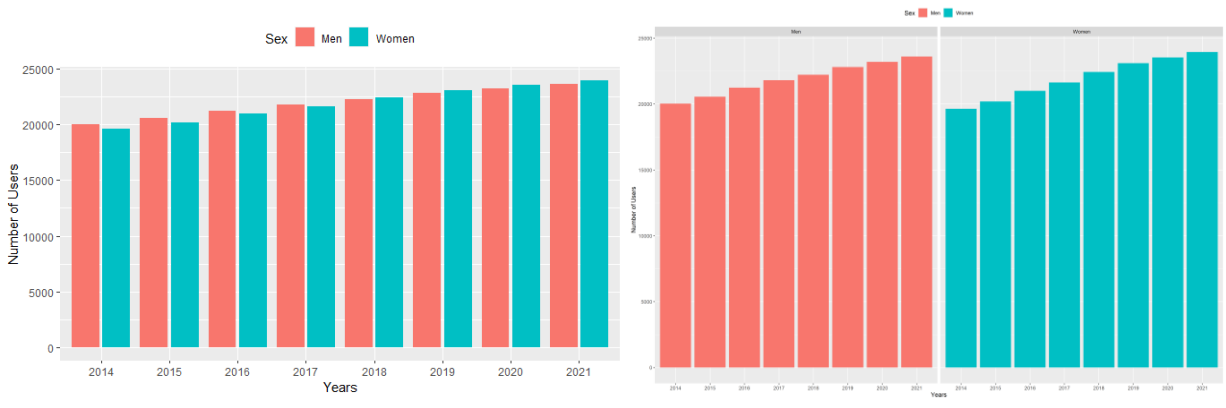

Figure 2.10. Bar graphs for Sex data

On further computation, the mean internet users of both men and women appears to be very similar and is further confirmed by a Pie chart.

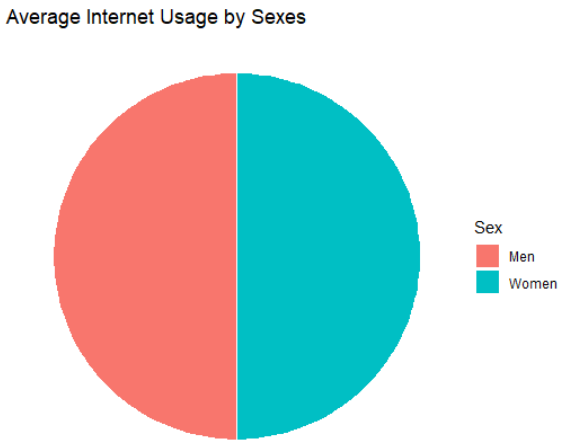

Figure 2.11. Pie graphs for Sex data

Dataset 3: Age and Sex Data.

This data consists of 2 categorical variables and a numerical variable. The bar plots are divided based on age-groups. We can observe again that there is not much difference in the usage between men and women.

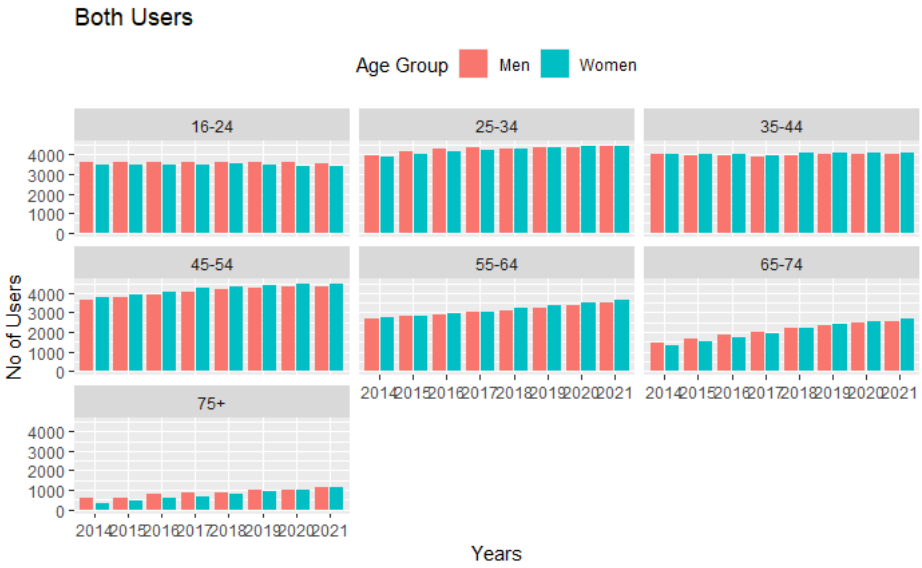

Figure 2.12. Bar chart for men and women users in each age – group.

A box plot shows that internet usage among women appears to vary more with each year. A large variation in usage is observed for both men and women in age group 65-74.

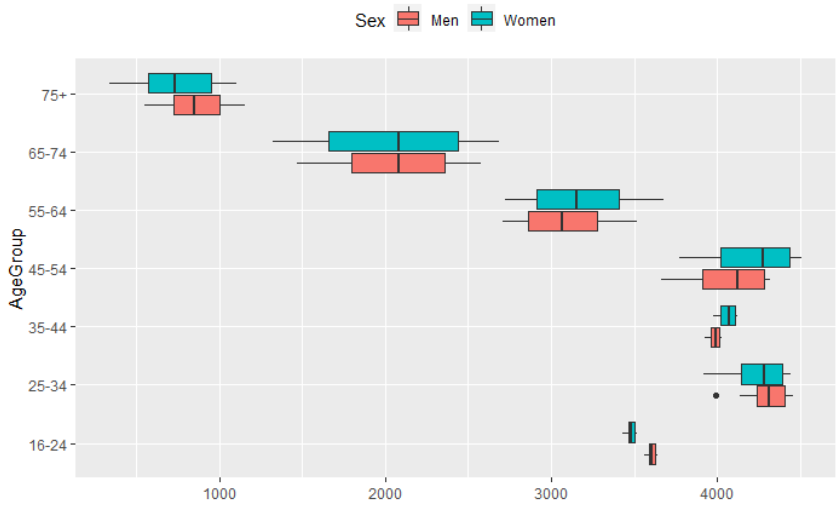

Figure 2.13. Box plot for men and women in each age – group.

The age groups of 35-44 and 16-24 appear to have a stable number of users with each year. Less variation could indicate that number of users in that age group can be predicted to not increase by much the next year. An outlier is noticed in age – group 25-34 for men, which indicates that there is a drop is usage in one of the years. An abnormal observation is known as an outlier.

Dataset 4: Age and Disability

On considering Disability category for all age groups combined, the number of "Not Equality Act disabled" users is much larger than its counterpart. This is displayed with clarity in the box plot.

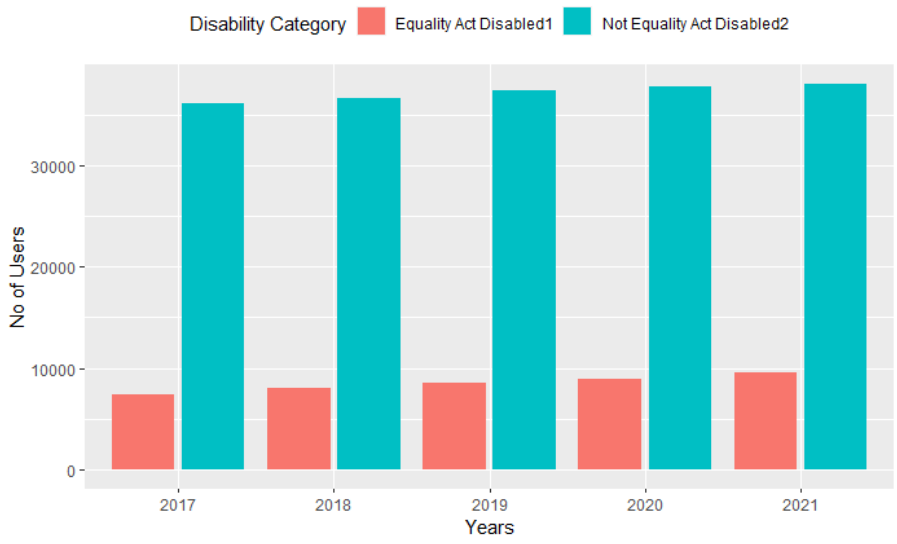

Figure 2.14. Bar graph for Disability data

The box plot shows large difference in how much the internet is used between the groups. This could also account for the number of people that are actually considered "Disabled" in the total population.

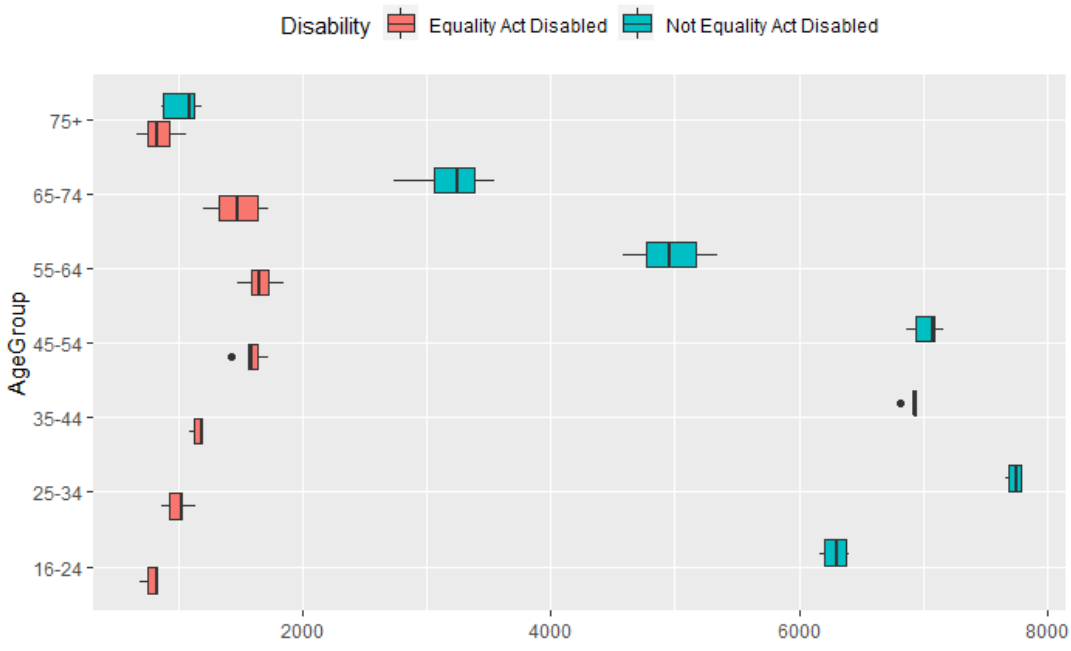

Figure 2.15. Box plot for Disability groups in each age-group

Dataset 5: Ethnicity Data

The ethnicity dataset can be represented with a simple bar plot which shows an overwhelming number of users belonging to “White” ethnicity. Other ethnicities in the dataset are “Indian”, “Bangladeshi”, “Chinese”, etc.

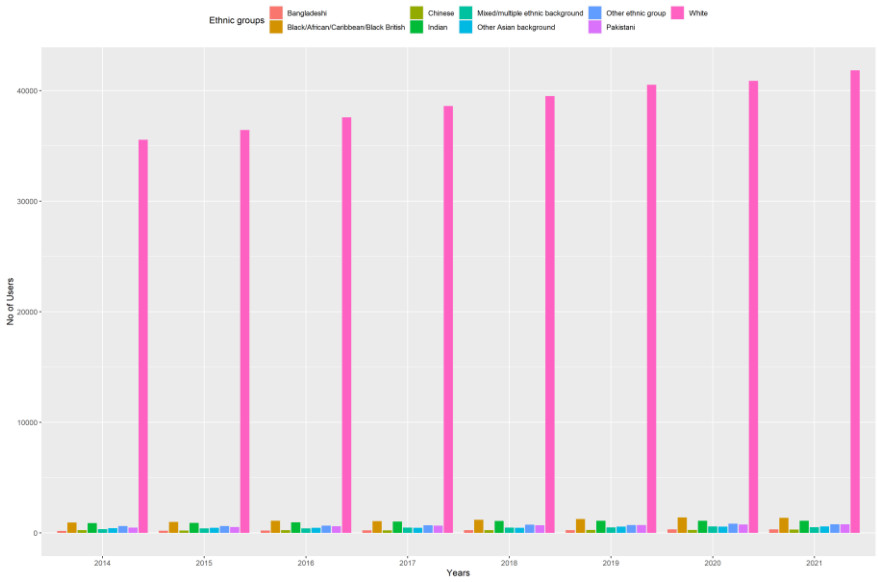

Figure 2.16. Bar plot of Ethnicity Dataset

Keeping objectives in mind, it is understood that the study should be focused on “White” ethnicity as the prospective clients of the company would most probably belong to this group.

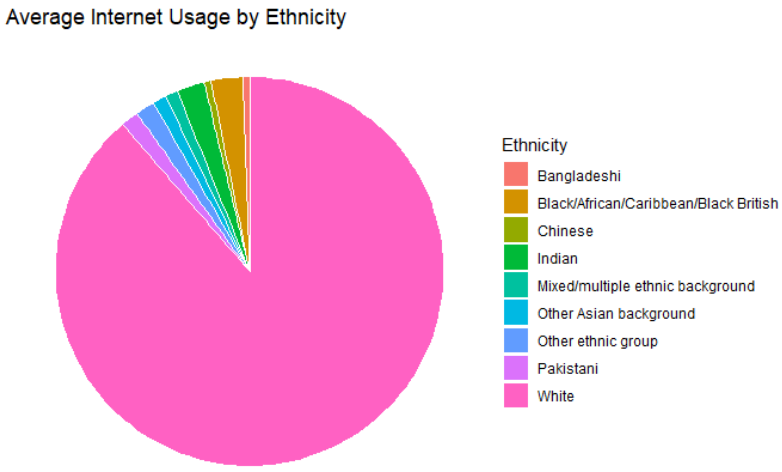

Figure 2.17. Pie graph of mean Ethnicity data shows majority “White” users.

Dataset 6: Economic Activity Data

The ethnicity dataset clearly shows us that “Employees” dominate the statistics of this dataset. They are key users in this category.

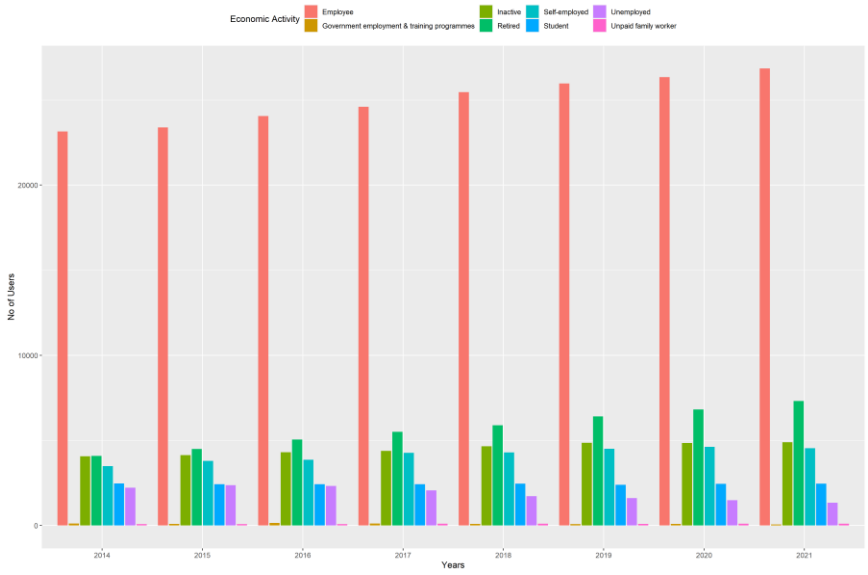

Figure 2.18. Bar Plot of Economic Activity Dataset

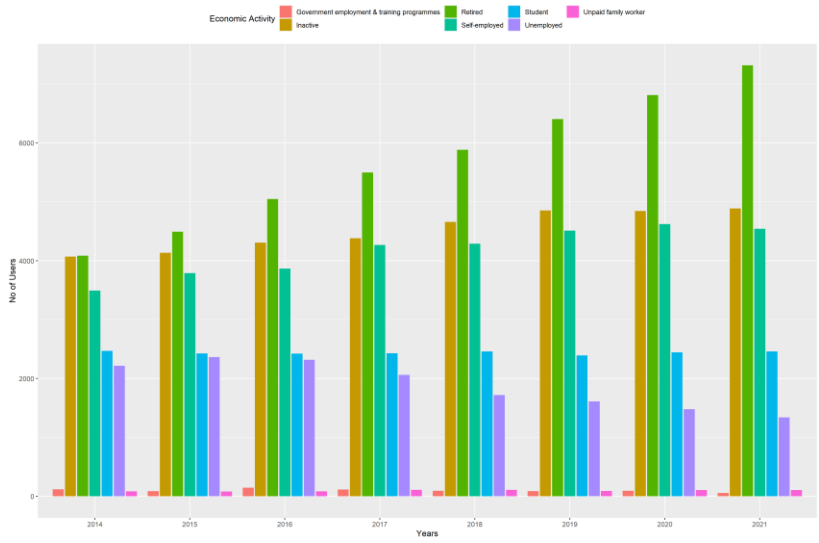

Figure 2.19. Bar plot of all economic activity data without “Employee”

Besides “Employee” group, there appear to be other groups with a rising usage of internet. With the highest being “Retired” group, “Inactive” and “Self-employed” close after.

Analysis

Dataset 1:

Earlier we inferred that some age groups appear to have more affect on the number of internet users. Hence, we test to find results for the objective of finalizing age-groups that play an important role in acquiring customers as a new company.

Defining the Hypotheses.

We first aim to answer the question "Is there a difference in usage among age – groups over the given period of time, 8 years (2014 – 2021)?". It also answers the questions about the relationship between the number of internet users and the age - group.

H_0 = There is no difference or no relationship.

H_1= There is a difference or there exists a relationship between the variables.

Hence:

X = List of age – groups

Y = number of internet users in each

We apply ANOVA test on the data with $\alpha = 0.05$, and obtain p = <2e-16 for $F_{6,49}$. Therefore, because the p< 0.05 we reject the Null hypothesis, H_0. This indicates that there exists a strong relationship between Age – Group and the Number of Users.

Regression Analysis.

To further study the effects that the categorical variable has on the quantitative data, we split the dataset and apply linear regression on each level of the age – group variable. We will eliminate the age – groups with the lowest number of users. Hence, here we only consider the groups 16 – 24, 25 – 34, 35 – 44, 45 – 54, and 55 – 64.

1. Age Group 16 – 24:

   The distribution in this data appears to be nearly normal but skewed left. However, the estimated slope is negative, indicating that with each year, the number of internet users is decreasing in this age – group. The Multiple R – squared value indicates that the variation in number of users can be explained by Years by 32.48% which is quite low. Also, considering $\alpha = 0.05$, p > 0.05, we can say that this model is not suited for the data of Age Group 16 – 24.

```
Call:
lm(formula = Years ~ Users, data = age_data1)

Residuals:
   Min     1Q Median     3Q    Max
-2.469 -1.856  0.152  1.473  2.560

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2195.22823  104.60898  20.985 7.63e-07 ***
Users         -0.02508    0.01476  -1.699     0.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.174 on 6 degrees of freedom
Multiple R-squared:  0.3248,    Adjusted R-squared:  0.2123
F-statistic: 2.887 on 1 and 6 DF,  p-value: 0.1402
```

```
Call:
lm(formula = Years ~ Users, data = age_data2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4232 -0.2905  0.1379  0.5728  0.9867

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.960e+03  8.623e+00 227.252  4.9e-13 ***
Users       6.795e-03  1.011e-03   6.722 0.000527 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9059 on 6 degrees of freedom
Multiple R-squared:  0.8828,    Adjusted R-squared:  0.8632
F-statistic: 45.18 on 1 and 6 DF,  p-value: 0.0005273
```

Figure 3.1(a) and (b). Results of Linear Regression on age-groups 16-24 and 25-34.

2. Age Group 25 – 34:

   The distribution in this data appears to be nearly normal but skewed right. The Multiple R – squared value indicates that the variation in number of users can be explained by Years by 88.28% which is quite significant. Also, p<0.05 and is nearly 0 so we conclude that the model is significant for this age – group.

3. Age Group 35 – 44:

   Since p = 0.1175 > 0.05, we understand that the model is not suited to test this data.

```
Call:
lm(formula = Years ~ Users, data = age_data3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9791 -0.3781  0.2437  1.4422  1.8631

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.886e+03  7.182e+01  26.265 2.01e-07 ***
Users       1.631e-02  8.927e-03   1.827    0.117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.121 on 6 degrees of freedom
Multiple R-squared:  0.3574,    Adjusted R-squared:  0.2503
F-statistic: 3.337 on 1 and 6 DF,  p-value: 0.1175
```

Figure 3.2. Results for age – group 35 – 44.

4. Age Group 45 – 54:

   The results signify, that this age – group has very significant coefficients and hence, the 3 asterisks. We can also notice a high R – squared value and p<0.05. This data can be used to predict growth in internet usage.

```
Call:
lm(formula = Years ~ Users, data = age_data4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55713 -0.40231 -0.07819  0.21867  1.04782

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.980e+03  3.499e+00  565.85 2.06e-15 ***
Users       4.571e-03  4.219e-04   10.83 3.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5835 on 6 degrees of freedom
Multiple R-squared:  0.9514,    Adjusted R-squared:  0.9433
F-statistic: 117.4 on 1 and 6 DF,  p-value: 3.664e-05
```

Figure 3.3. Results for age – group 45 – 54.

5. Age Group 55 – 64:

   This age group has one of the largest multiple R – squared values. The p value is also extremely close to 0. The model is the best fit for the data of this age – group.

```
Call:
lm(formula = Years ~ Users, data = age_data5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30220 -0.08564  0.02092  0.11162  0.23856

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.993e+03  7.503e-01 2656.07  < 2e-16 ***
Users       3.928e-03  1.196e-04   32.85 5.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1967 on 6 degrees of freedom
Multiple R-squared:  0.9945,    Adjusted R-squared:  0.9936
F-statistic:  1079 on 1 and 6 DF,  p-value: 5.29e-08
```

Figure 3.4. Results for age – group 55 – 64.

Dataset 2:

Through the visualization of the data we inferred that there is not much difference in the usage between men and women. Now, we will statistically test this inference.

Defining the Hypotheses.

The question we wish to answer is "Is there a difference in internet usage between men and women in the UK?".

H_0 = There is no difference.

H_1= There is a difference.

Hence:

X = sex category (levels: "men" and "women")

Y = number of internet users in each

We apply T-test on this data because the categorical variable X contains only 2 levels. We test at 95% confidence, so $\alpha = 0.05$, and obtain p = 0.9964 for $T_{13.361}$. Therefore, because the p > 0.05 we accept the Null hypothesis, H_0. This indicates that this category has no relevant effect on the number of users in the sample data. Hence, we do not concentrate on this data anymore as it does not help us reach the objective.

Dataset 3:

The Age and Sex data can be used to deduce if both categories are significant in determining the number of users.

Defining the Hypotheses.

The question we wish to answer is "Is there a difference in internet usage between men and women of different age – groups in the UK?".

H_0 = There is no difference.

H_1= There is a difference.

Hence:

X = sex category (levels: "men" and "women") and Age Groups

Y = number of internet users in each

We apply ANOVA on this data and get,

```
             Df   Sum Sq  Mean Sq F value Pr(>F)
AgeGroup      6 1.58e+08 26330968   409.3 <2e-16 ***
Sex           1 7.00e+00        7     0.0  0.992
Residuals   104 6.69e+06    64331
---
```

Figure 3.5. Results of ANOVA on dataset 3

From this, we can observe that Age-Group plays the major factor in the number of people that use the internet. As previously confirmed, Sex is not a factor that can be considered relevant for this project.

Therefore, we accept the Null hypothesis only when Age – Group is the variable, while we reject the null hypothesis when Sex is considered as X.

Regression Analysis.

We get similar results on applying Linear Regression on this dataset. We can see the asterisk signs that mark significant variables or factors are highlighted only for Age – Groups variable.

```
Call:
lm(formula = Users ~ AgeGroup + Sex, data = agesex_pivot)

Residuals:
    Min      1Q  Median      3Q     Max
-726.94 -109.08   -0.56  133.69  630.06

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3543.19      67.79  52.269  < 2e-16 ***
AgeGroup25-34   718.75      89.67   8.015 1.71e-12 ***
AgeGroup35-44   478.88      89.67   5.340 5.49e-07 ***
AgeGroup45-54   595.25      89.67   6.638 1.48e-09 ***
AgeGroup55-64  -419.38      89.67  -4.677 8.80e-06 ***
AgeGroup65-74 -1489.75      89.67 -16.613  < 2e-16 ***
AgeGroup75+   -2747.81      89.67 -30.642  < 2e-16 ***
SexWomen          0.50      47.93   0.010    0.992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 253.6 on 104 degrees of freedom
Multiple R-squared:  0.9594,    Adjusted R-squared:  0.9566
F-statistic: 350.8 on 7 and 104 DF,  p-value: < 2.2e-16
```

Figure 3.6. Results of Linear Regression on Dataset 3

Dataset 4:

This data has 2 categorical variables with one categorical variable having 2 levels.

Defining the Hypotheses.

The question we wish to answer is "Does the number of internet users differ with existence of a disability?".

H_0 = There is no difference.

H_1= There is a difference.

Hence:

X = disability category (levels: "Equality Act Disabled" and "Not Equality Act Disabled")

Y = number of internet users in each

We apply T-test on this data because the categorical variable X contains only 2 levels. We test at 95% confidence, so $\alpha = 0.05$, and obtain p = 1.276e-11. Therefore, because the p < 0.05 we reject the Null hypothesis, H_0. This was already visible to us with Visualized data.

Dataset 5:

Defining the Hypotheses.

The question we wish to answer is "Does ethnicity have an effect on the number of users?".

H_0 = There is no effect.

H_1= There is an effect.

Hence:

X = ethnicity category

Y = number of internet users in each

We apply ANOVA on this data because the categorical variable X contains more than 2 levels. We get $F_{7,56}$ Therefore, because the p = <2e-16 which is less than 0.05 we reject the Null hypothesis, H_0. This was already visible to us with Visualized data. We noticed a difference in the user levels between ethnic groups.

The ethnicity data contained one Ethnic group, "White" with very high values while other has number of users below 50% of the total. We do not focus on these groups. We focus on the majority group of "White" to maximize profits.

Regression Analysis.

```
Call:
lm(formula = Years ~ Users, data = ethnic_white)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32339 -0.20119  0.01647  0.18701  0.28323

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.975e+03  1.683e+00 1173.72  < 2e-16 ***
Users       1.096e-03  4.324e-05   25.34 2.49e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2545 on 6 degrees of freedom
Multiple R-squared:  0.9907,    Adjusted R-squared:  0.9892
F-statistic: 642.4 on 1 and 6 DF,  p-value: 2.485e-07
```

Figure 3.5. Results for ethnic group "White".

To further confirm our analysis, we apply linear regression on the data for "White" ethnic group. We see the distribution of residuals is skewed right. We have a very high Multiple R – squared value, indicative that this model is the best fit for the data. The slope is positive hence increasing number of users in this group.

Dataset 6:

Economic activity does appear to have a variation in number of users with each group in the visualization of the data.

Defining the Hypotheses.

The question we wish to answer is "Does economic activity have an effect on the number of users?".

H_0 = There is no effect.

H_1= There is an effect.

Hence:

X = economic activity category

Y = number of internet users

We apply ANOVA on this data because there are multiple economic activities. We get $F_{6,49}$Therefore, because the p = <2e-16 which is less than 0.05 we reject the Null hypothesis, H_0.

Regression Analysis.

After considering the groups with larger usage of the internet, the analysis has been narrowed down for groups "Employee", "Self – Employed", "Retired", "Student" and "Inactive".

1. Employee

```
Call:
lm(formula = Years ~ Users, data = eco_emp)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3531 -0.2552  0.1073  0.1771  0.2896

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.974e+03  1.857e+00 1062.60  < 2e-16 ***
Users       1.757e-03  7.426e-05   23.66 3.74e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2724 on 6 degrees of freedom
Multiple R-squared:  0.9894,    Adjusted R-squared:  0.9876
F-statistic: 559.9 on 1 and 6 DF,  p-value: 3.739e-07
```

Figure 3.6. Results of Linear Regression for Employee

2. Self – employed

```
Call:
lm(formula = Years ~ Users, data = eco_2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0387 -0.3383 -0.1161  0.2848  1.3781

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.994e+03  2.947e+00 676.458 7.04e-16 ***
Users       5.716e-03  7.035e-04   8.125 0.000187 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7637 on 6 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9028
F-statistic: 66.01 on 1 and 6 DF,  p-value: 0.0001867
```

Figure 3.7. Results of Linear Regression for Self-Employed

3. Retired

```
Call:
lm(formula = Years ~ Users, data = eco_3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09998 -0.05137 -0.01844  0.07457  0.10158

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 2.005e+03  1.597e-01 12557.37  < 2e-16 ***
Users       2.169e-03  2.758e-05    78.64 2.85e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08237 on 6 degrees of freedom
Multiple R-squared:  0.999,     Adjusted R-squared:  0.9989
F-statistic:  6184 on 1 and 6 DF,  p-value: 2.847e-10
```

Figure 3.8. Results of Linear Regression for Retired

4. Student

We notice a decreasing slope in this group. Hence, it is not favourable to focus on this group as potential customers.

```
Call:
lm(formula = Years ~ Users, data = eco_4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3153 -1.8314  0.0375  1.5610  3.6225

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.031e+03  9.410e+01  21.586 6.45e-07 ***
Users       -5.663e-03  3.861e-02  -0.147    0.888
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.641 on 6 degrees of freedom
Multiple R-squared:  0.003573,  Adjusted R-squared:  -0.1625
F-statistic: 0.02151 on 1 and 6 DF,  p-value: 0.8882
```

Figure 3.9. Results of Linear Regression for Student

5. Inactive

```
Call:
lm(formula = Years ~ Users, data = eco_5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89894 -0.36406  0.08774  0.28343  0.85939

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.985e+03  3.081e+00  644.44 9.42e-16 ***
Users       7.108e-03  6.803e-04   10.45 4.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6039 on 6 degrees of freedom
Multiple R-squared:  0.9479,    Adjusted R-squared:  0.9392
F-statistic: 109.2 on 1 and 6 DF,  p-value: 4.509e-05
```

Figure 3.10. Results of Linear Regression for Inactive

Conclusion

The project is designed to find the target groups that will bring forward profit for the company in a new venture. Hence, it is highly important to keep in mind groups that have high internet usage, which directly correlates with increased profit.

After completing the Analysis through various statistical methods, we can confidently focus on the target groups which are:

1. Age Groups:
    a. 25 – 34
    b. 45 – 54
    c. 55 – 64
2. Ethnicity:
    a. White
3. Economic Activity:
    a. Employee
    b. Self – Employed
    c. Retired
    d. Inactive

References

https://www.learnbymarketing.com/tutorials/linear-regression-in-r/

http://www.stat.cmu.edu/~brian/701/notes/paper-structure.pdf

https://whatis.techtarget.com/definition/statistical-analysis

https://www.scribbr.com/category/statistics/

https://www.investopedia.com/terms/t/t-test.asp

https://www.datacamp.com/community/tutorials/linear-regression-R

https://statisticsbyjim.com/hypothesis-testing/hypothesis-tests-significance-levels-alpha-p-values/

Practical Statistics for Data Scientists by Peter Bruce, Andrew Bruce, and Peter Gedeck.

https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3