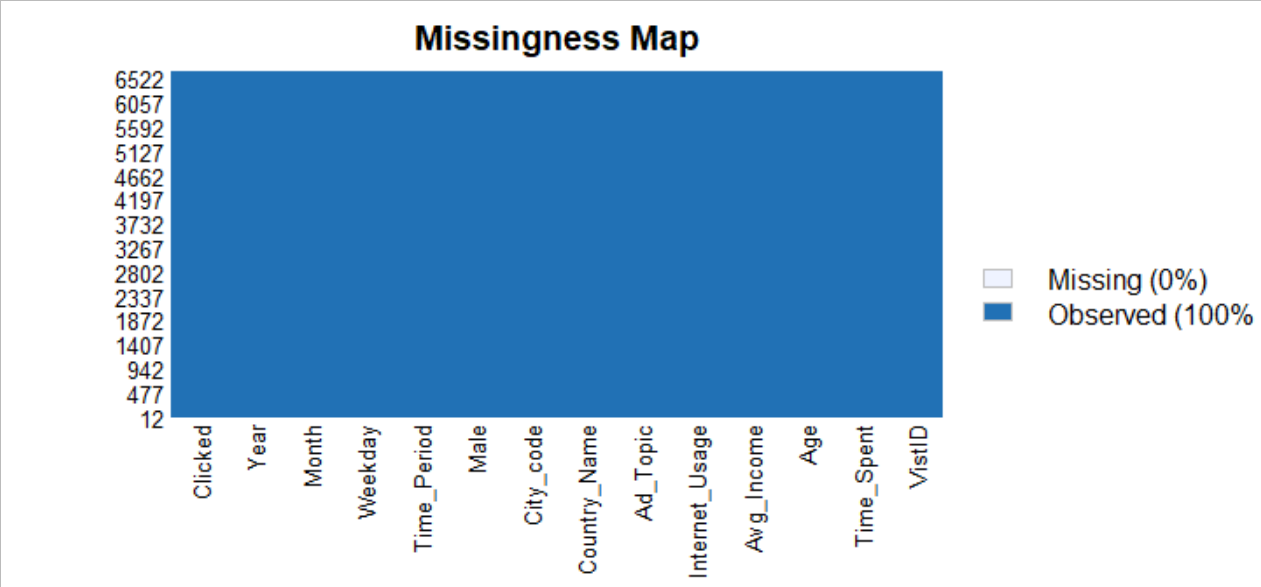
## **Problem description:**

The goal of the case study is to Predict who is likely going to click on the Advertisement so it can contribute to the more revenue generation to the organization.

**Data Wrangling**

1. The dataset consists of 6657 rows and 14 columns.
2. 4 numerical columns and 10 categorical columns.
3. Clicked (whether the Ad is clicked or not)– dependent variable
4. Ad\_Topic, Country\_Name, City\_code, Male, Time\_Period, Weekday, Month, Clicked are converted into factor variables.
5. No missing value is found using Amelia package.



1. No outliers are found in the continuous variables.
2. Data summarization:

Continuous Variable

|  |  |  |  |
| --- | --- | --- | --- |
| **Time\_Spent** | **Age** | **Avg\_Income** | **Internet\_Usage** |
| Min. :32.60 | Min. :19.00 | Min. :13996 | Min. :104.8 |
| 1st Qu.:55.20 | 1st Qu.:28.00 | 1st Qu.:48454 | 1st Qu.:145.7 |
| Median :70.66 | Median :36.00 | Median :58183 | Median :193.6 |
| Mean :66.85 | Mean :37.26 | Mean :55930 | Mean :184.9 |
| 3rd Qu.:79.57 | 3rd Qu.:46.00 | 3rd Qu.:65957 | 3rd Qu.:222.3 |
| Max. :91.43 | Max. :61.00 | Max. :79485 | Max. :270.0 |

Categorical Variable

|  |  |  |  |
| --- | --- | --- | --- |
| **Ad\_Topic** |  |  |  |
| product\_1: 238 | product\_18: 234 | product\_26: 227 | product\_8:226 |
| product\_10: 232 | product\_19: 195 | product\_27:250 | product\_9:216 |
| product\_11:216 | product\_2: 144 | product\_28:229 |  |
| product\_12:227 | product\_20: 234 | product\_29:219 |  |
| product\_13: 247 | product\_21:207 | product\_3:112 |  |
| product\_14:226 | product\_22:266 | product\_30:234 |  |
| product\_15: 210 | product\_23:224 | product\_4: 215 |  |
| product\_16: 224 | product\_24: 230 | product\_5:216 |  |
| product\_17 : 237 | product\_25: 255 | product\_7:223 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Country\_Name** | **City\_code** | **Time\_Period** | **Weekday** |
| Canada : 46 | City\_1 :2559 | Early-Morning: 674 | Friday :986 |
| Saudi Arabia : 44 | City\_2 :1675 | Evening :1266 | Monday :981 |
| Eritrea : 43 | City\_3 :1075 | Mid-Night :1145 | Saturday :931 |
| Venezuela : 43 | City\_4 : 673 | Morning :1228 | Sunday :979 |
| Benin : 42 | City\_5 : 364 | Night :1140 | Thursday :926 |
| Papua New Guinea: 42 | City\_6 : 195 | Noon :1204 | Tuesday :935 |
| (Other) :6397 | City\_7 : 81 |  | Wednesday:919 |
|  | City\_8 : 31 |  |  |
|  | City\_9 : 4 |  |  |

|  |  |  |
| --- | --- | --- |
| **Month** | **Male** | **Clicked** |
| April :974 | No :3571 | 0:3619 |
| February:944 | Yes:3086 | 1:3038 |
| January :988 |  |  |
| July :930 |  |  |
| June :941 |  |  |
| March :930 |  |  |
| May :950 |  |  |

1. **Information values** for numerical and categorical variables

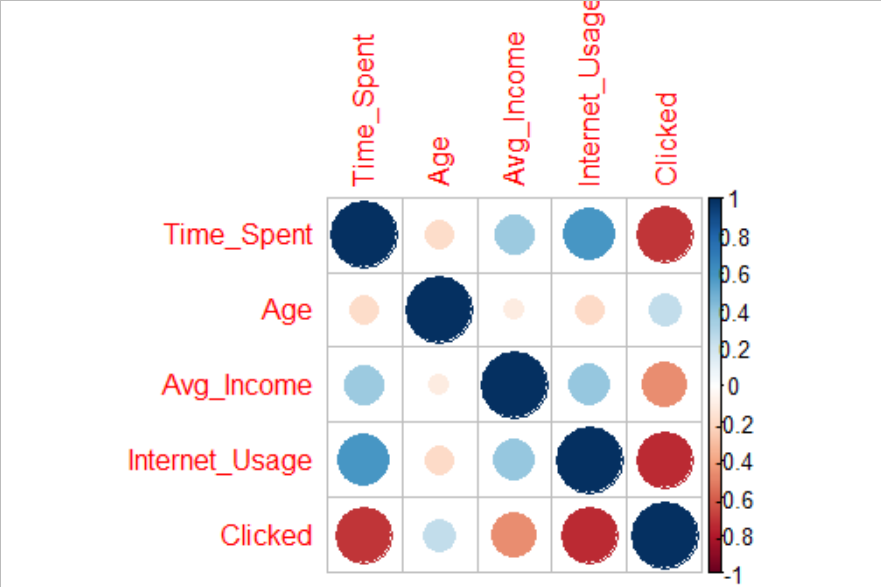
|  |  |
| --- | --- |
| **Variable name** | **IV** |
| VistID | 0.00122691688105001 |
| Time\_Spent | 2.52143422985652 |
| Age | 0.383865956481418 |
| Avg\_Income | 1.1984247899787 |
| Internet\_Usage | 2.19799193751923 |
| Year | 0 |
| Ad\_Topic | 0.13425 |
| Country\_Name | 0.260874 |
| City\_code | 0.270947 |
| Male | 0.003032 |
| Time\_Period | 0.144382 |
| Weekday | 0.002215 |
| Month | 0.003634 |

1. There are 237 different unique countries in our dataset and no single country is too dominant. A large number of unique elements will not allow a machine learning model to establish easily valuable relationships. For that reason, this variable will be excluded.

Year contains only one value that is 2020 (year in which the dat was collected), it will not be of any use in the model building.

**Country, VistID, Year** are unnecessary columns, hence have been removed.

**Exploratory Data Analysis**

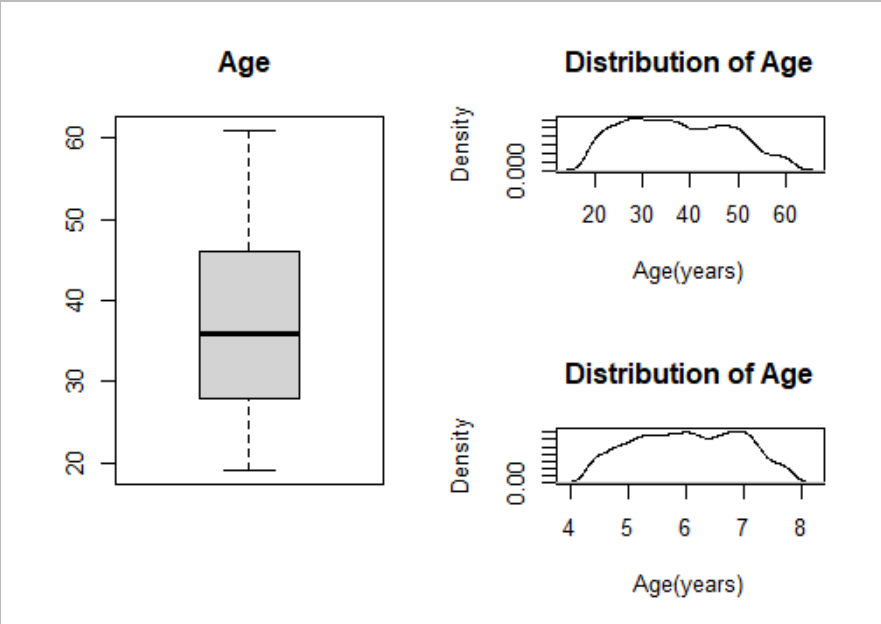


Time\_Spent is positively correlated with Internet\_usage and negatively correlated with Clicked.

Which means the more time people spent on the internet, the less likely are they to click on the Ad.

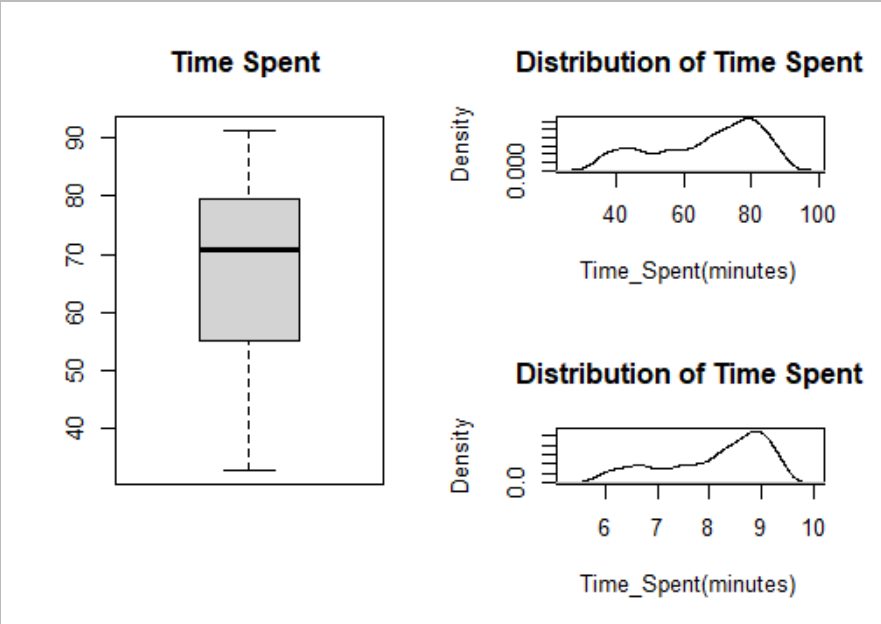
Clicked is negatively correlated with Time\_Spent and Internet\_Usage.

1. Boxplot and distribution of Age.



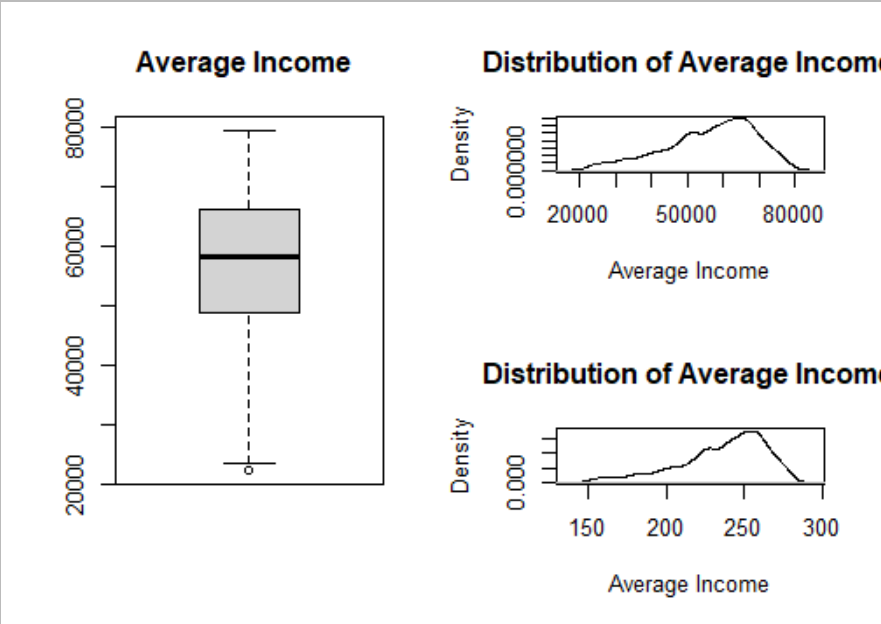
**Observation findings:**

1. No outliers in Age.
2. Median Age of people visiting the website = 36 years.
3. Maximum distribution is observed between the age of 20 and 45 years after which the distribution is dropping till 60 years of age.
4. This implies that users of the age between 20 and 45 years can be the main target group for the marketing campaign. Hypothetically, if we have a product intended for middle-aged people, this is the right site for advertising. Conversely, if we have a product intended for people over the age of 60, it would be a mistake to advertise on this site.



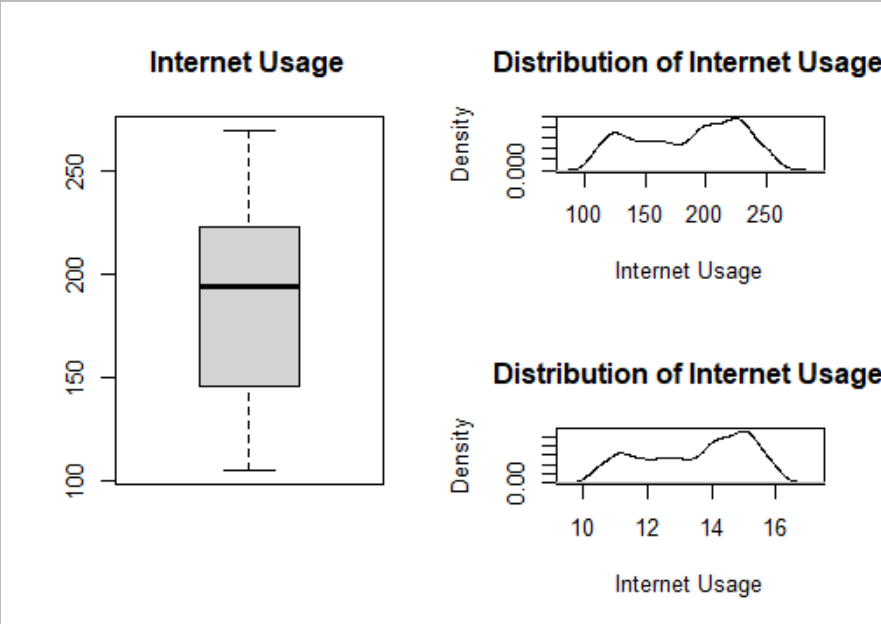
**Observation findings:**

1. No outliers in Time\_Spent.
2. Median Time\_Spent by people visiting the website = 70 years.
3. Maximum distribution is observed between the age of 70 and 80 mintures after which the distribution is dropping till 100 minutes.
4. This implies that users are most likely to spend 60-80minutes on the site.



**Observation findings:**

1. No outliers in Avg\_Income.
2. Median Average Income of people visiting the website = 58183.
3. Maximum distribution is observed between the income of 50000 and 80000.
4. An interesting fact from the table is that the smallest area income is 22206 and the highest is 79,485. This means that site visitors are people belonging to different social classes. It can also be concluded that we are analyzing a popular website since users spend between 32 and 91 minutes on the website in one session. These are really big numbers.



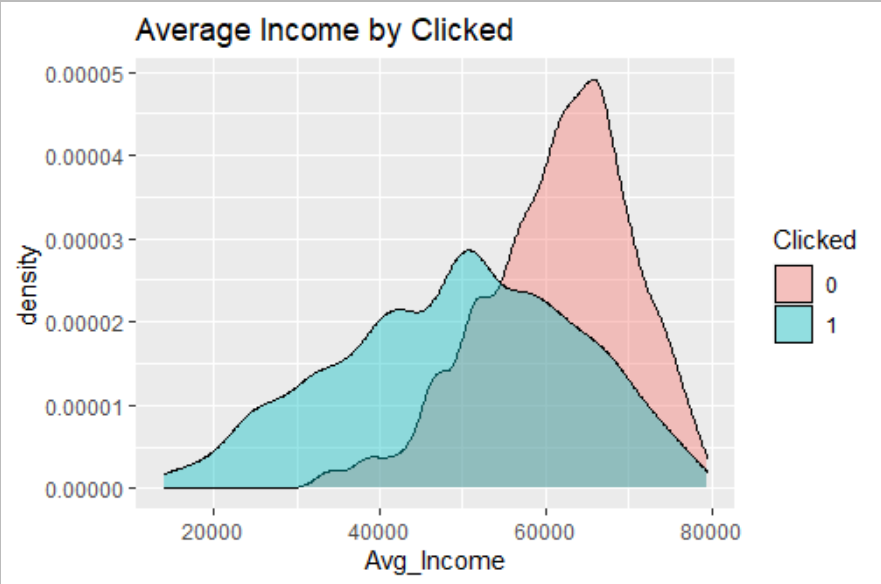
**Observation findings:**

1. No outliers in Internet\_usage.
2. Median Internet usage of people visiting the website = 193.8.
3. Maximum distribution is observed between of 200 to 250 minutes.

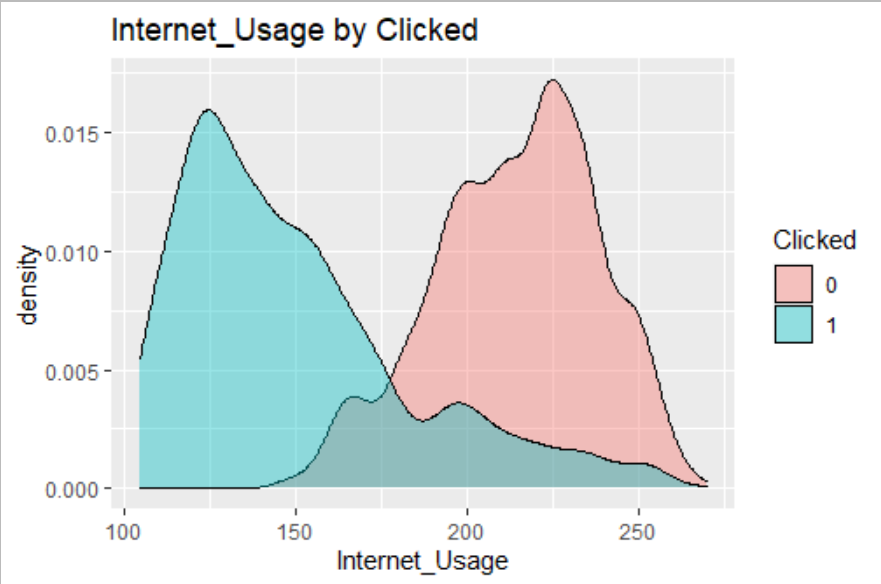
From the figure above, it is clear that users who spend more time on the internet also spend more time on the site.



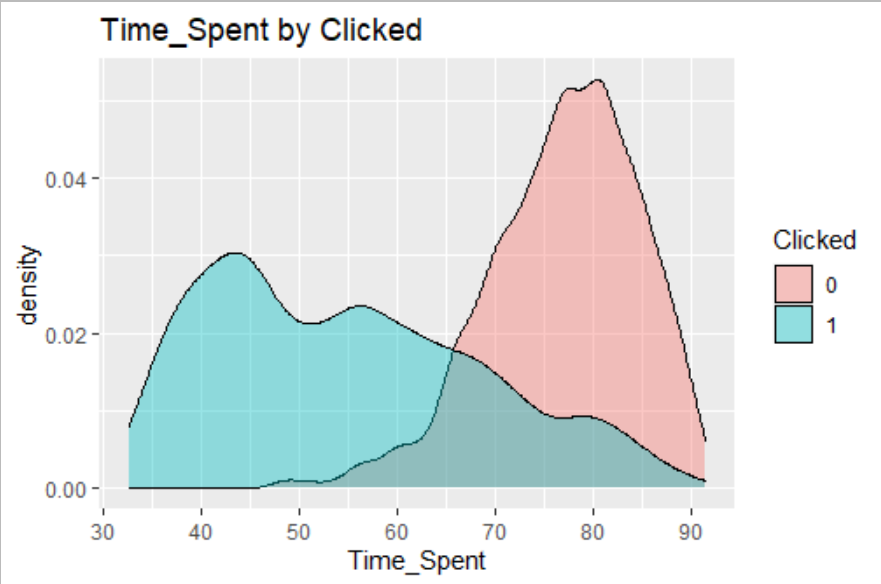
People who **clicked on the ad have a median income of 58000** compared to people who do not click on the ad who have a median income of above 60000.



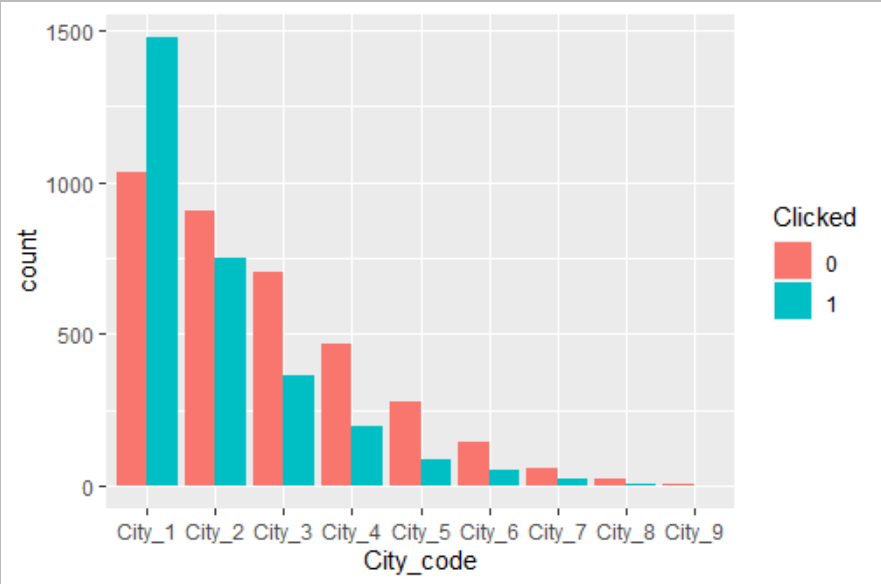
1. **Relationship of Clicked with other continuous variables.**



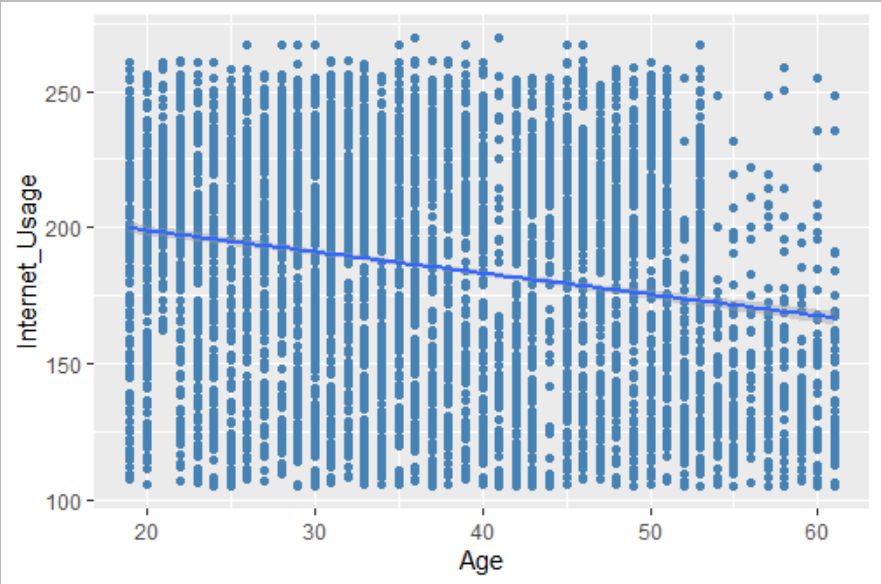
People who are **likely** to click on the ad **spend 100-150 minutes** on the internet compared to people **who spend more than 200 minutes** (they are less likely to click on the ad).



People who are **likely to click on the ad spend 30-65 minutes** on the site, people who spend more than **60 minutes are less** likely to click on the ad.



People belonging to **City\_1** clicked on the Ad more followed by City\_2.

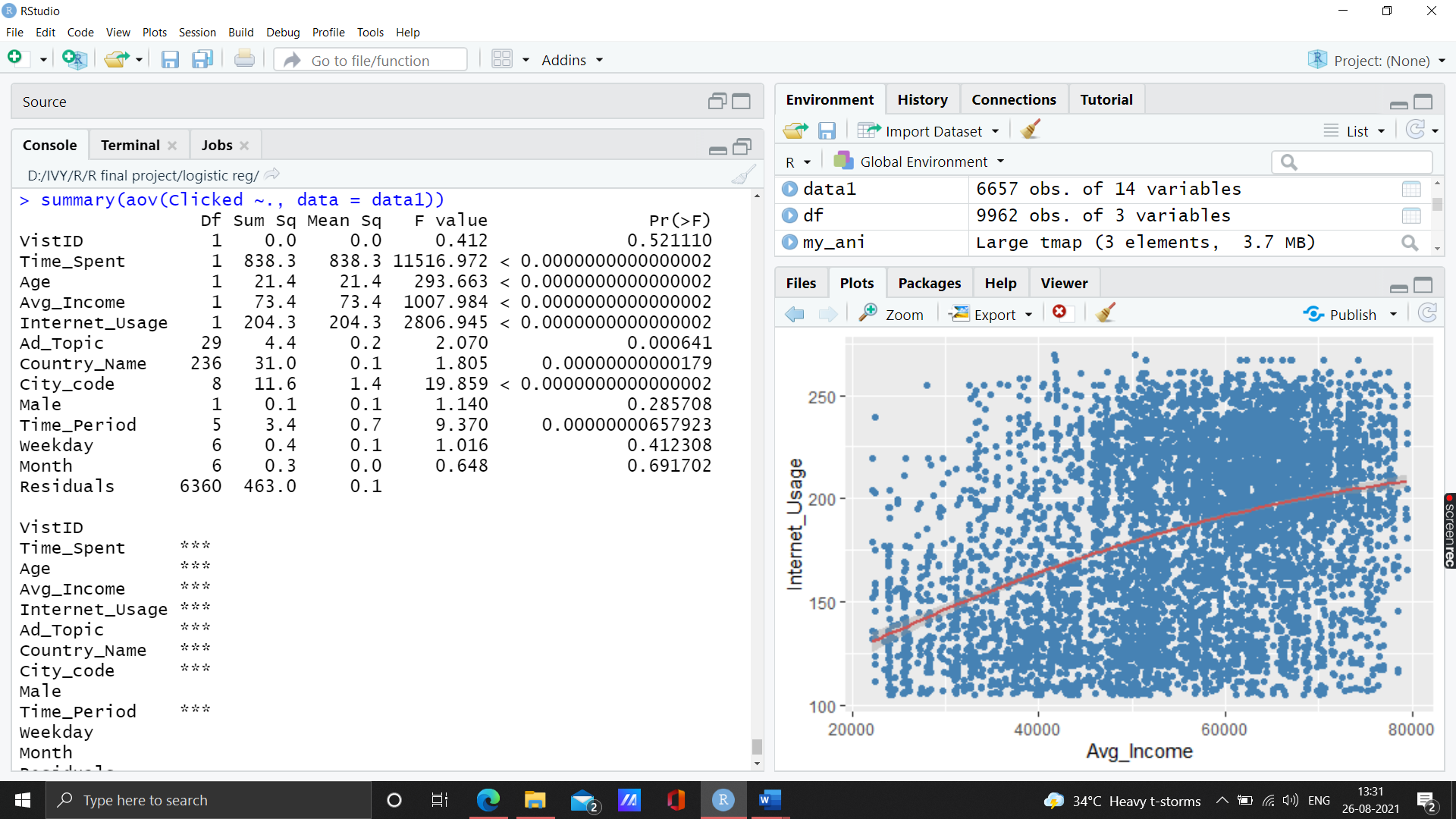


**Internet\_usage and Age are negatively correlated**, with the age increasing people get less time to browse the internet.



**Internet\_Usage and Avg\_income are postitvely correlated**. With increasing income it is seen that people are more likely to increase their internet usage.

1. **Anova** test



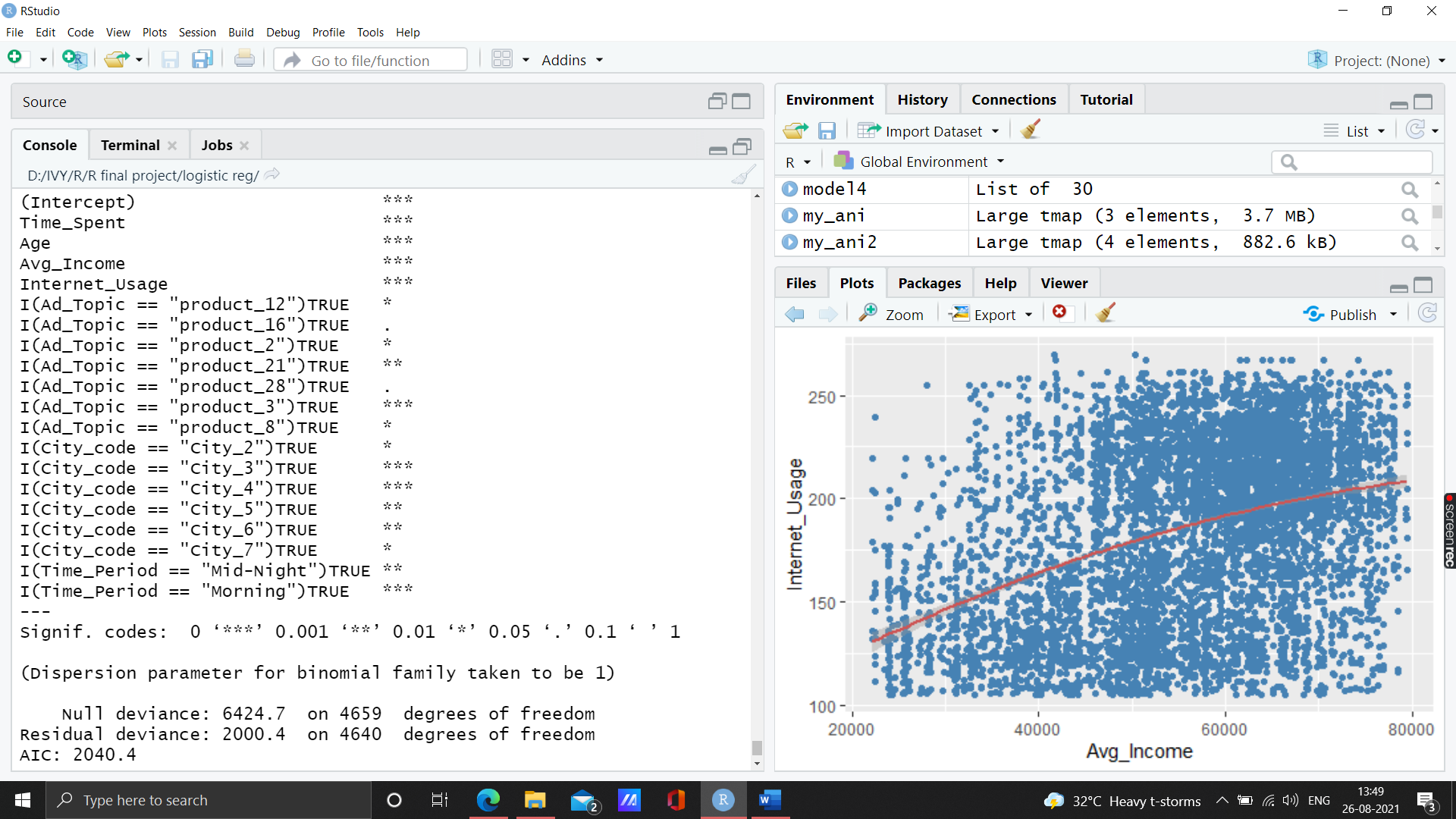
Variables **Time\_Spent, Age, Avg\_Income, Internet\_Usage, Ad\_Topic, Country\_Name ,City\_code and Time\_Period** are significant based on the p-value.

1. **Model Building and prediction**

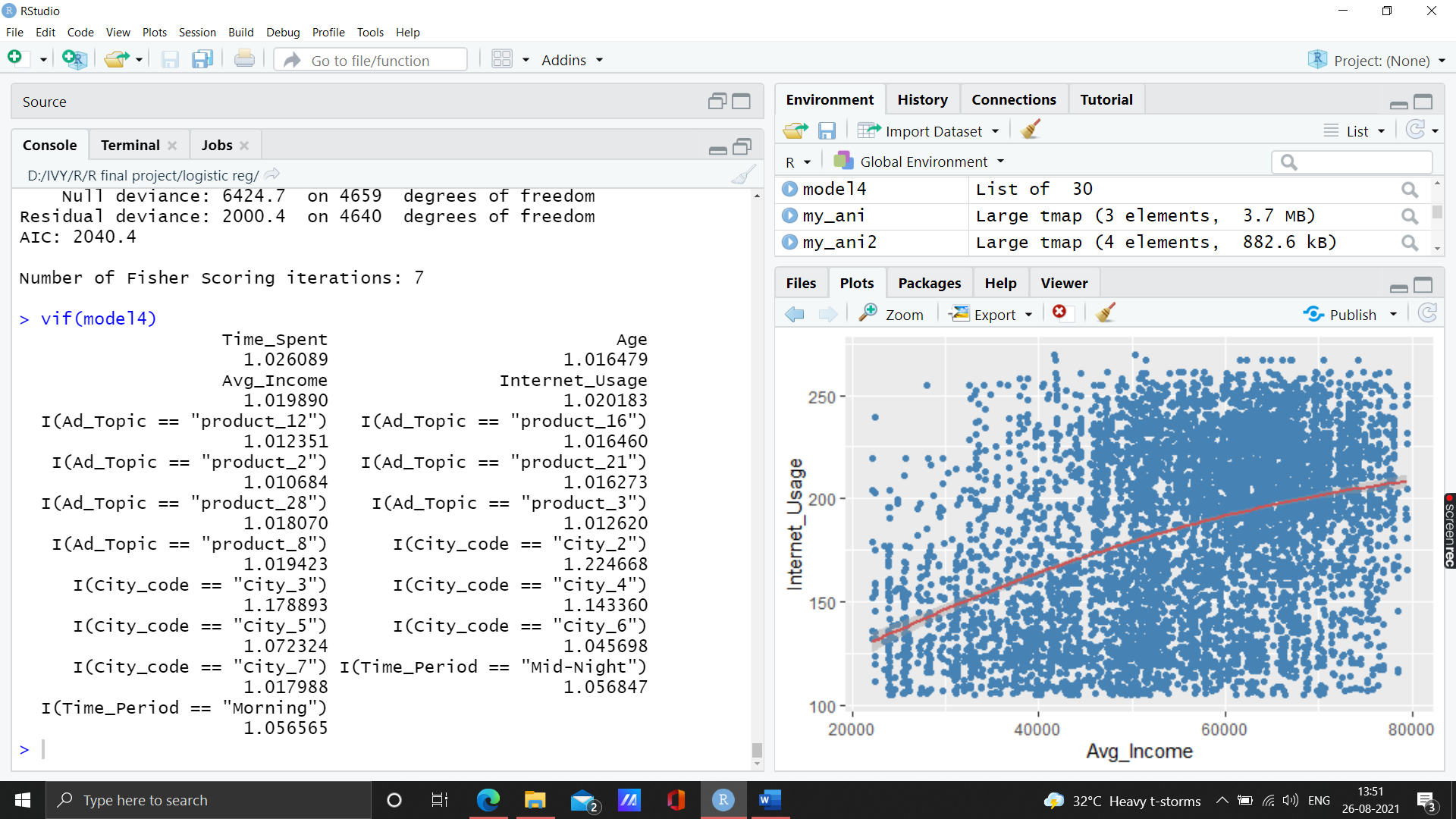
The dataset have been divided into 70% training set and 30% test data set.

4660 observations, 11 columns are in training set.

1997 observations, 11 columns are in test dataset.



After training the model we have **19 significant variables** at 90% confidence interval.



In the Variance Inflation test we find that, All the variable have Variance Inflation Factor less than <2 and p-value <0.0001, which shows that there is no significant correlation exists between the variables. Hence, **no multicollinearity exists**.

1. **Model Diagnostics**: Checking the overall fitness of the model
2. **Wald Test:**

Chi-squared test:

X2 = 1204.6, df = 20, P(> X2) = 0.0

Since, p-value is less then 0.001, hence we reject Ho(null hypothesis) that the all Bi=0.

1. **Lagrange Multiplier or Score Test** (Assess whether the current variable significantly improves the model fit or not)

*Lagrange Multiplier or Score Test* (Assess whether the current variable significantly improves the model fit or not).

***Null Hypothesis***: Null model = Baseline Model

***Alternate Hypothesis***: Null model <> Baseline Model

Null model will explain the 'y' less and actual model will explain the 'y' more

If p-value is less than 0.05 then we reject the null hypothesis that the model is no better than chance.

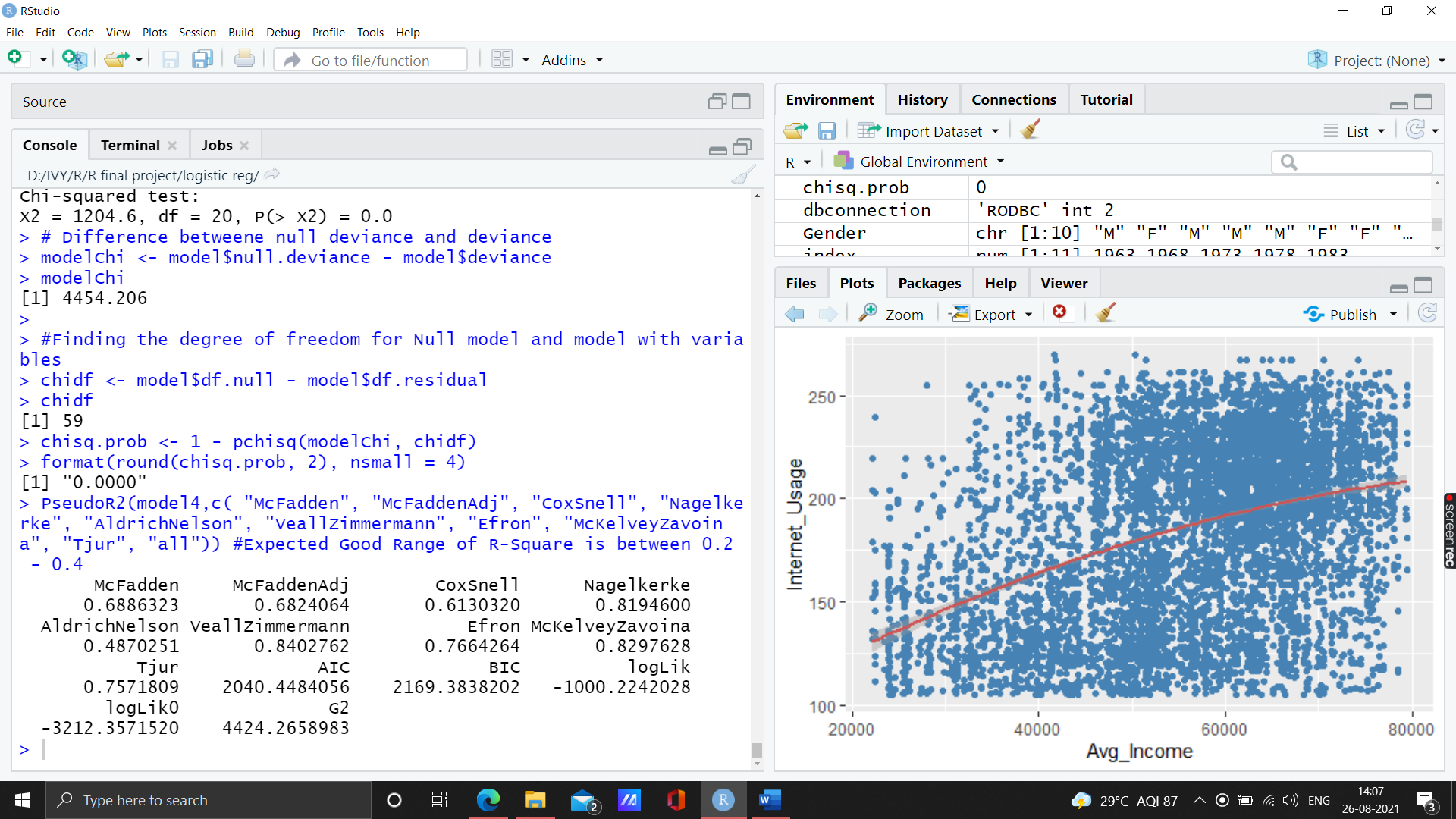
*chisq.prob <- 1 - pchisq(modelChi, chidf)*

*> format(round(chisq.prob, 2), nsmall = 4)*

*"0.0000"*

If p value is less than 0.05 then we reject the null hypothesis that the model is no better than chance.

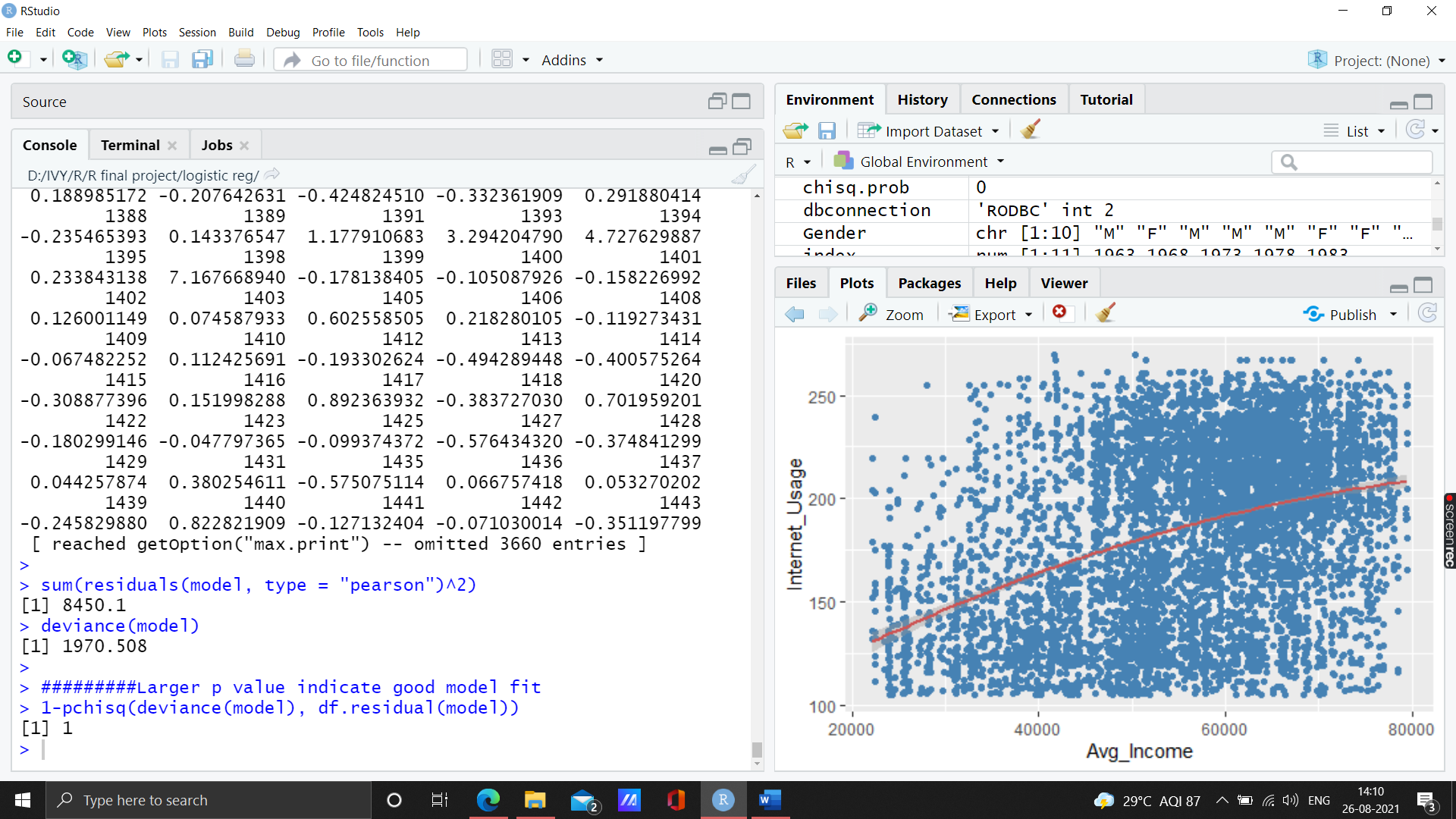
1. **Predicting power of the model using R2**



Expected Good Range of R-Square is between 0.2 - 0.4

1. **Lackfit Deviance** for assessing whether the model where

#Ho (Null Hypothesis): Observed Frequencies/probabilties =Expected Frequencies/probabilities



Larger p value indicate good model fit

Thus, we accept the Null Hypothesis Ho that Observed Frequencies = Expected Frequencies.

1. **Metrics- Fit Statistics**

**Metric Values**

1. threshold 0.5681591

2. specificity 0.8857546

3. sensitivity 0.9707856

4. AUC 0.9614605

5. AccuracyRate 0.9319742

6. Gini 0.9229210



ROC curve

**AUC(Area under the curve : 96% > 80%**

Good model fit.

The higher the area under the ROC curve, the better the prediction ability.

1. **Confusion Matrix**

**Actual**

**Predicted** 0 1

0 **2459** 243

1 74 **1884**

**True negative: 2459**

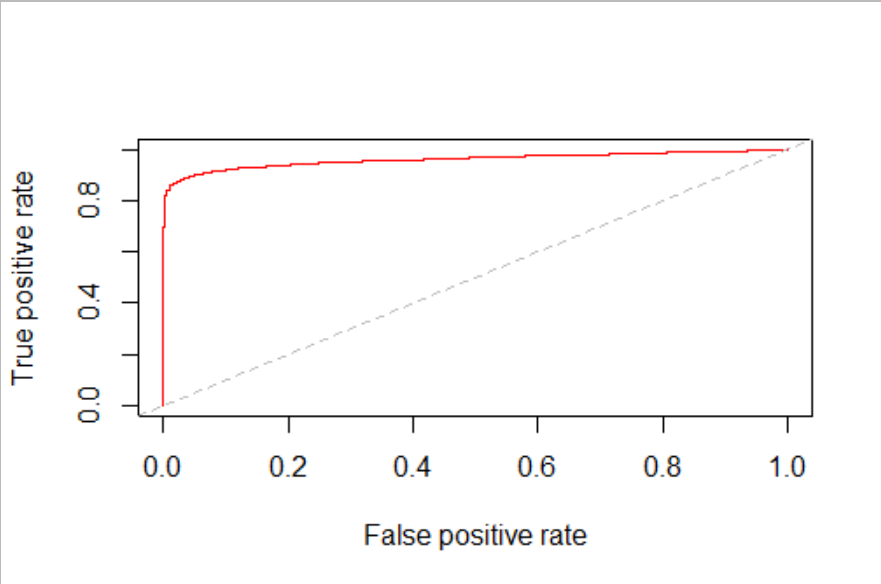
**False Negative: 1884**

1. **KS Statistics Calculation**

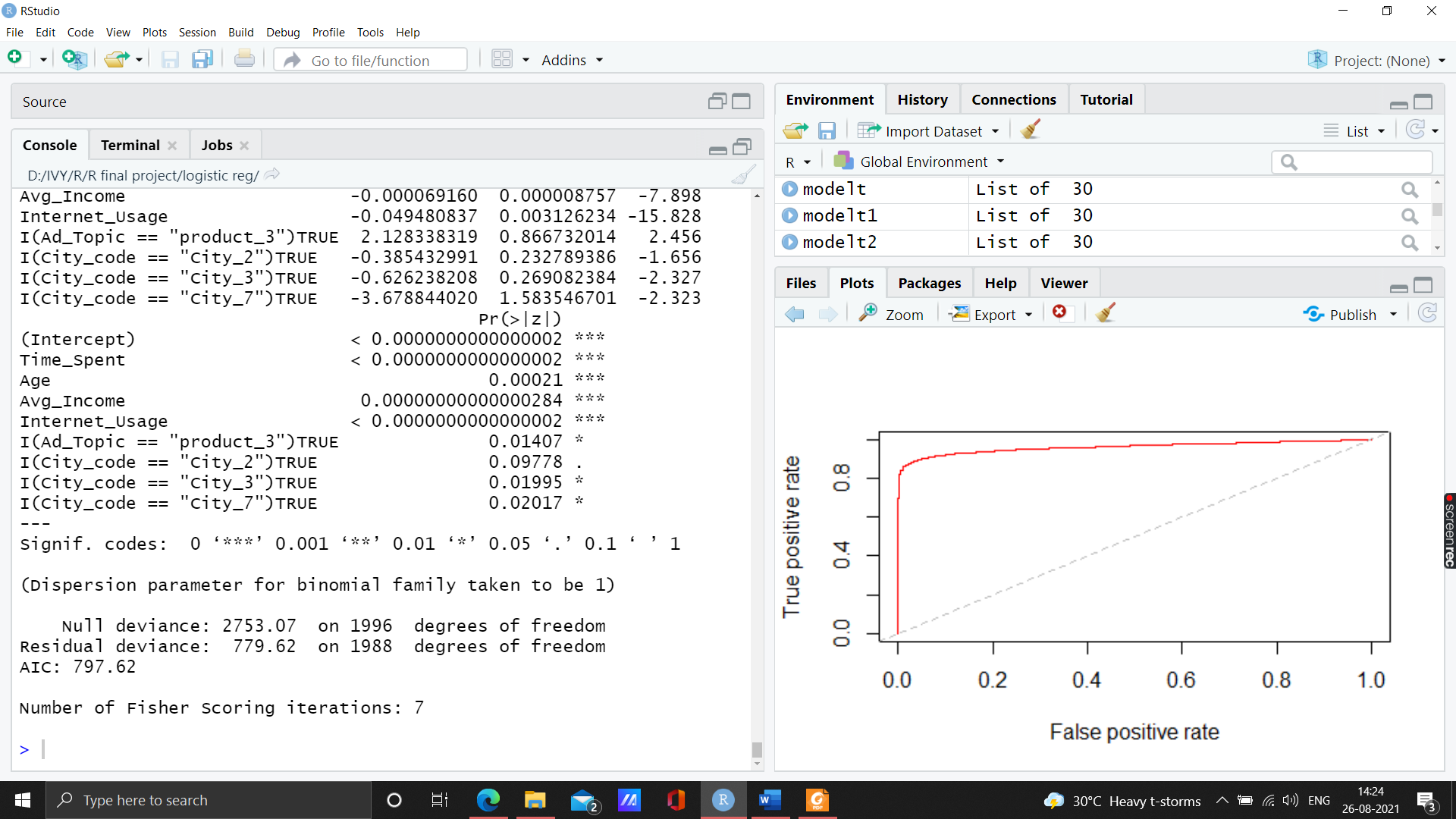
ks1.logit

Thumb rule : should lie between 0.4 - 0.7

**0.8565042**



1. Model has been build, now testing the model on the test dataset



After using the training model on the test dataset, we find that only 8 predictors namely, **Time\_Spent, Age, Avg\_Income, Internet\_Usage, I(Ad\_Topic == "product\_3"), I(City\_code == "City\_2"), I(City\_code == "City\_3")** and **I(City\_code == "City\_7")** are significant.

Vif test

**Time\_Spent Age**

**1.034660 1.018247**

**Avg\_Income Internet\_Usage**

**1.035215 1.027789**

**I(Ad\_Topic == "product\_3") I(City\_code == "City\_2")**

**1.006241 1.088122**

**I(City\_code == "City\_3") I(City\_code == "City\_7")**

**1.070797 1.021793**

In the Variance Inflation test we find that, All the variable have Variance Inflation Factor less than <1.7 and p-value <0.0001, which shows that there is no significant correlation exists between the variables. Hence, **no multicollinearity exists**.

1. **Model Diagnostics**: Checking the overall fitness of the model
2. **Wald Test:**

Chi-squared test:

X2 = 496.1, df = 9, P(> X2) = 0.0

Since, p-value is less then 0.001, hence we reject Ho(null hypothesis) that the all Bi=0.

1. **Lagrange Multiplier or Score Test** (Assess whether the current variable significantly improves the model fit or not)

*chisq.prob <- 1 - pchisq(modelChi, chidf)*

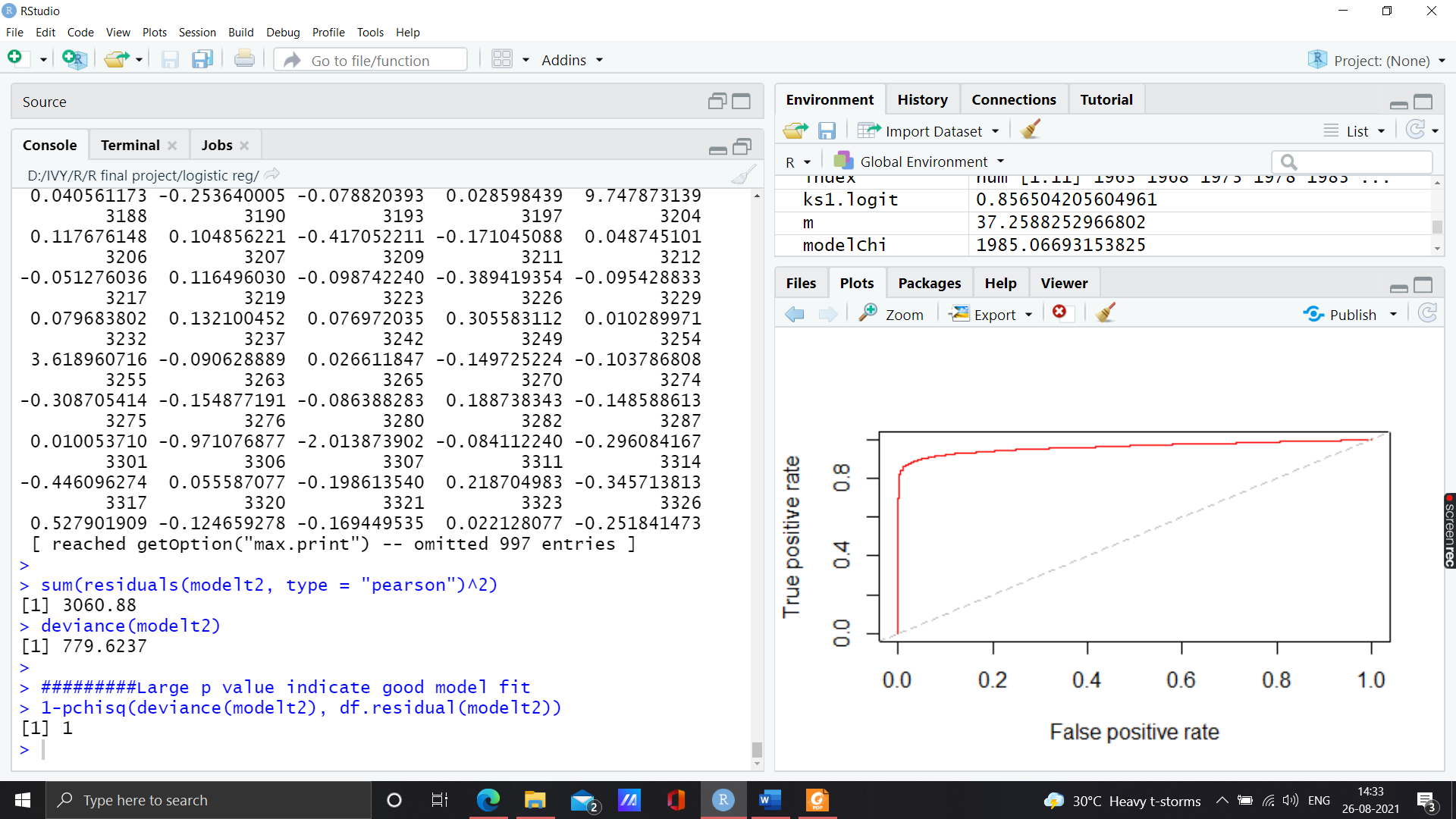
*> format(round(chisq.prob, 2), nsmall = 5)*

*"0.0000"*

If p value is less than 0.05 then we reject the null hypothesis that the model is no better than chance.

1. **Lackfit Deviance** for assessing whether the model where

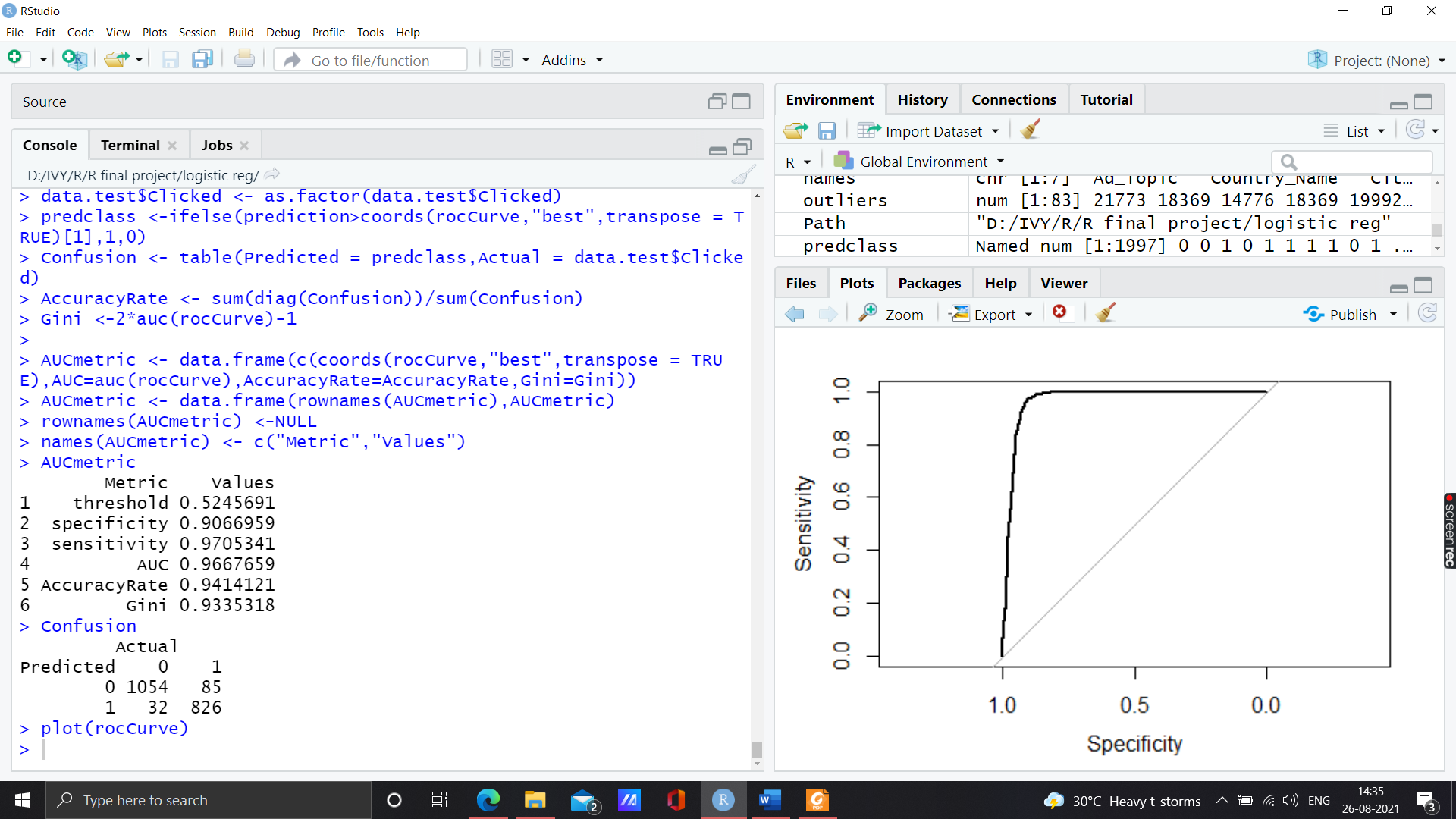
#Ho (Null Hypothesis): Observed Frequencies/probabilties =Expected Frequencies/probabilities

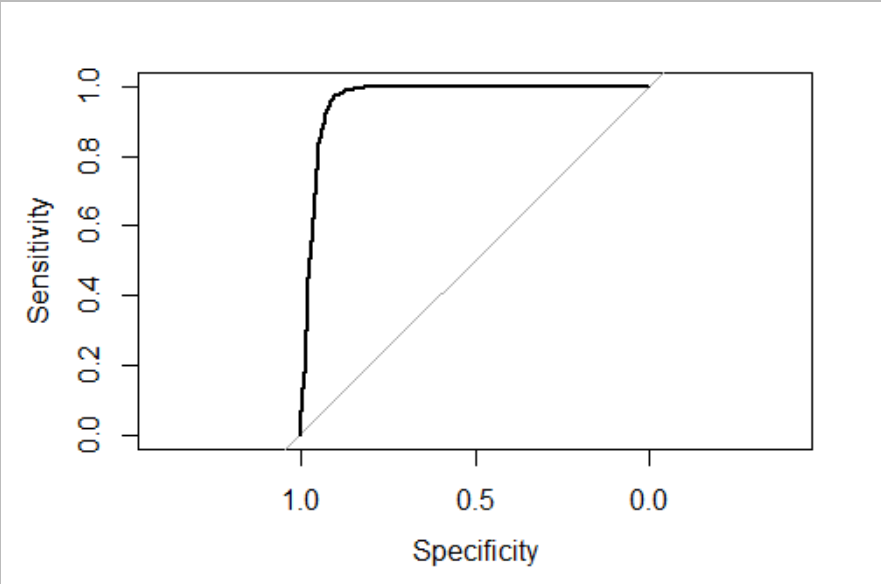


Larger p value indicate good model fit

Thus, we accept the Null Hypothesis Ho that Observed Frequencies = Expected Frequencies.

1. **Metrics- Fit Statistics**





ROC curve

**AUC(Area under the curve : 96.67% > 80%**

Good model fit.

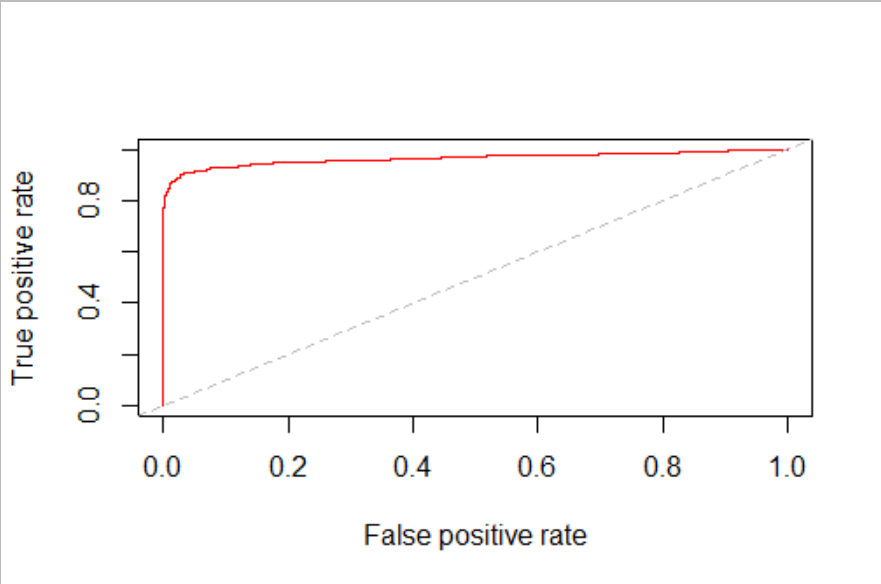
The higher the area under the ROC curve, the better the prediction ability.

1. **KS Statistics Calculation**

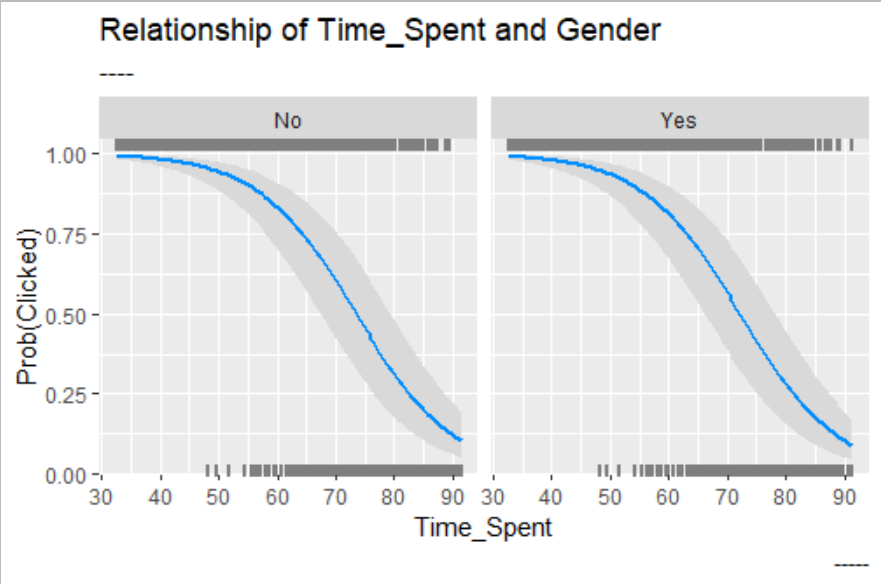
ks1.logit

Thumb rule : should lie between 0.4 - 0.7

**0.87723**

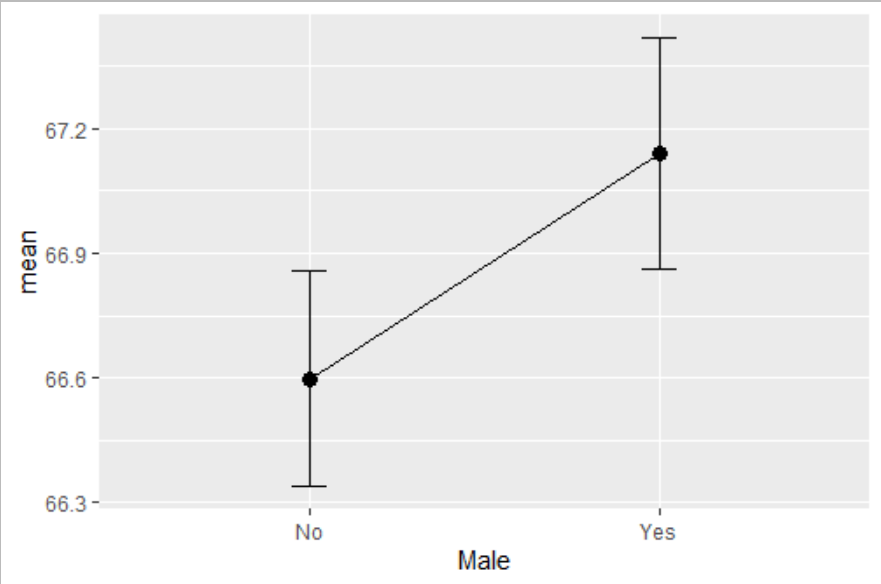


1. **Plots based on logistic regression model**



**Yes= Male, No= Female**

**Probability of clicking on the web add, of both male and female if the time spent on the website is 30- 55 minutes.**



calculating mean and standard errors by Gender and Time spent

Average time spent by females : 66.6 minutes

Average time spent by males : 67.1 minutes

# A tibble: 2 x 5

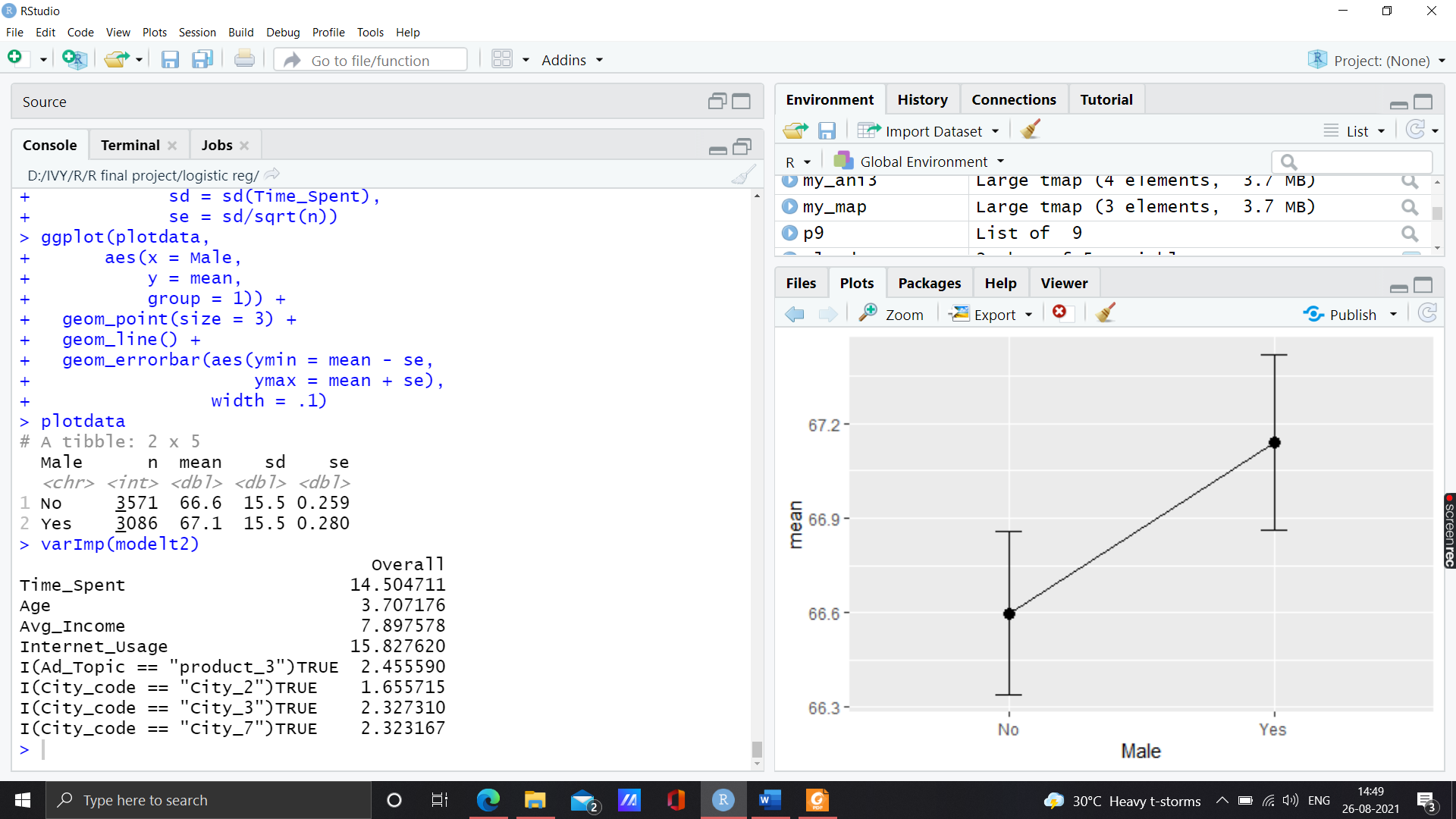
Male n mean sd se

<chr> <int> <dbl> <dbl> <dbl>

1 No 3571 66.6 15.5 0.259

2 Yes 3086 67.1 15.5 0.280

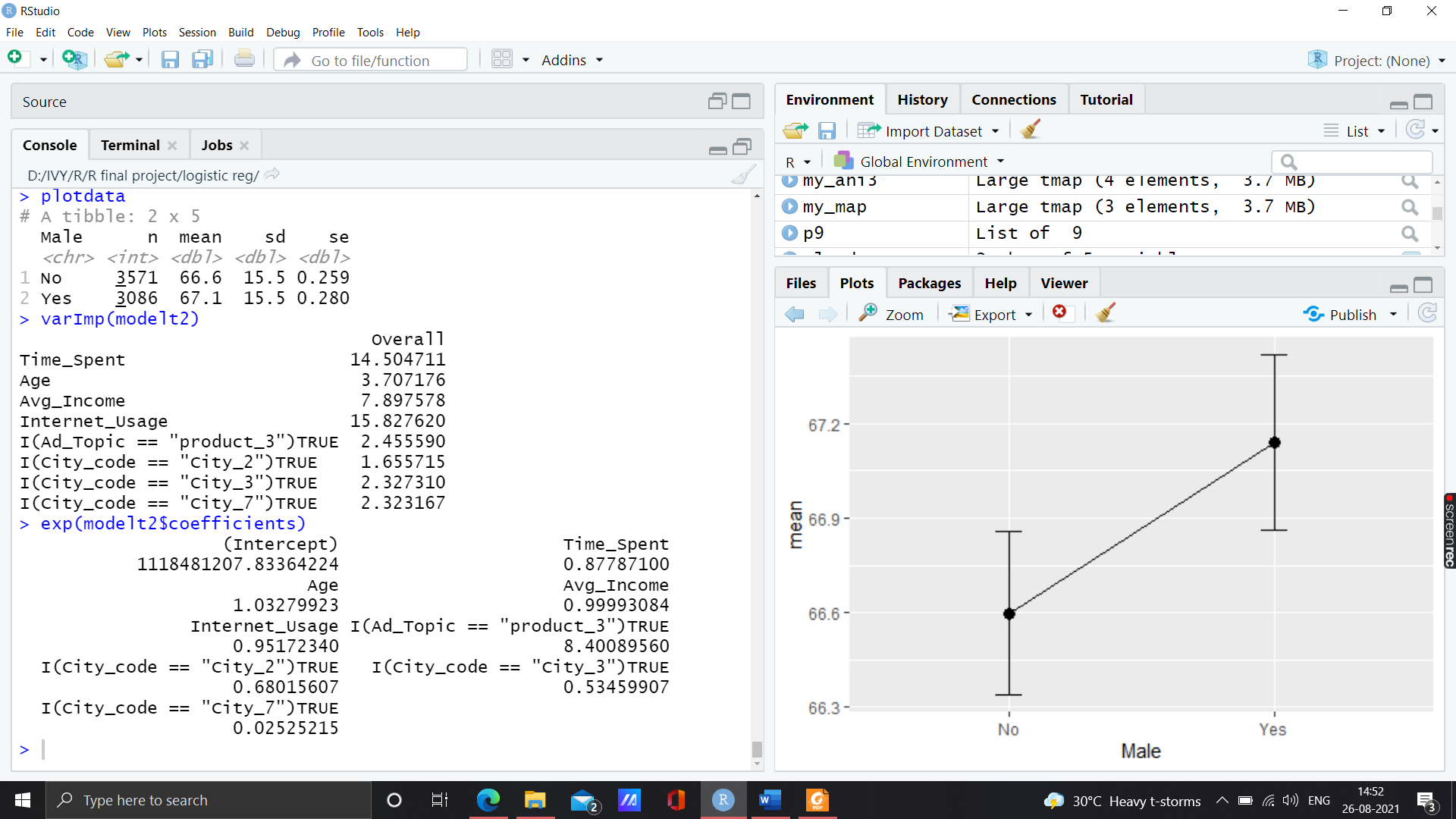
1. **Variable Importance:**



The most important variable: **Internet\_Usage followed by Time\_Spent.**

**Then comes Avg\_Income and Age**.

1. **Interpretation by co-efficients(odd ratio):**



1. **Age** : With the increasing age, a person is 1.0327 times likely to click on the web add.
2. **Time\_Spent**: If a person spends more time on the website then he is 0.87787 times likely to click on the Ad. Increase in time spent, less likely to click on the web add.
3. **Avg\_Income**: With the increase in average income, a person is 0.999 times likely to click on the web add. More Avg\_Income less likely to click on the web add.
4. **Internet\_Usage** : If a person uses internet for longer time on the website then he is 0.95172 times likely to click on the Ad. Increase in Internet\_Usage, less likely to click on the web add.
5. **I(Ad\_Topic== “product\_3” )**: If Ad\_Topic is product\_3 then the person is 8.40089 times more likely to click on the Ad.
6. **I(City\_code==”City\_2”):** If City\_code is City\_2 then the person is 0.6801 times likely to click on the Ad. This implies if the person is of City\_code : City\_2 then he is less likely to click on the web add.
7. **I(City\_code==”City\_3”)**: If City\_code is City\_3 then the person is 0.5346 times likely to click on the Ad. This implies if the person is of City\_code : City\_3 then he is less likely to click on the web add.
8. **I(City\_code==”City\_7”)**: If City\_code is City\_7 then the person is 0.0252 times likely to click on the Ad. This implies if the person is of City\_code : City\_7 then he is less likely to click on the web add.

|  |  |
| --- | --- |
| **Positive Impact on Ad\_Click** | **Negative Impact on Ad\_Click** |
| Age | Time\_Spent |
| I(Ad\_Topic== “product\_3” ) | Avg\_Income |
|  | Internet\_Usage |
|  | I(City\_code==”City\_2”) |
|  | I(City\_code==”City\_3”) |
|  | I(City\_code==”City\_7”) |