ML PROJECT WORK

Topic: Quantifying and Predicting a person's keenness towards work from home culture.

Group members:

Badrivishaal Shenoy

2048007

Angshuman Sinha

2048004

Contents-

Table of tables

Table of figures

1. INTRODUCTION

1.1. ABOUT THE DOMAIN: WORK FROM HOME CULTURE.

In this current pandemic scenario, we have seen that work from home has been promoted to keep ourselves safe; Now this work from home culture was something that always existed but people never opted for it as it made more sense to be at a place with a system where everyone must work in an orderly fashion, due to the pandemic which caused lock downs companies were forced to make their employees work from home, this has now been going on for more than a year.

1.2. OBJECTIVE:

Our goal from this analysis is to find out if we can predict a person's keenness towards the work from home culture, The positives, and negatives of this new formed culture.

1.3. PEOPLE THAT WOULD BENEFIT FROM THIS STUDY:

This study is conducted for more than just to classify/predict a person's keenness towards working from home, it is to see what factors effect a person's keenness towards working from home and to see if we can do anything to make this experience better. So, the people who would benefit the most are those businesses that provide home décor (for a more comfortable environment), service providers (WIFI, telecom) and most of all the people who are going to be working from home as this is a chance for them to asses this new culture and give reasons for their assessment.

2. ABOUT THE DATA

2.1 DATA COLLECTION

We had made a questionnaire with 30 questions regarding the topic of study and 166 responses were recorded. This questionnaire was sent to friends, family and acquaintances and based on their response our data was collected. The dimensions of dataset are 166 by 23.

2.2 DATASET DETAILS

Data Set Characteristics	Multivariate	**Number of Instances**	166
Feature Characteristics	Categorical/numeric	**Number of attributes**	23
Target Characteristics	Numeric 5-point feature	**Missing values?**	No
Associated task	Classification/Regression	**Domain**	Work from home culture

Table 1.1: Description about the dataset

2.3 FEATURE DESCRIPTION

1. Age in years –Age
2. The Gender - Gender
3. Are you an unmarried person, member of a nuclear family or a joint family? - famtype
4. Is your family dependent only on your income? - dependence
5. On a scale of 1 to 5, how would you rate work from home culture? - wfh_rating [Work from Home Culture rating]
6. What factors according to you will impact the work from home culture? - wfh_factor [factor]
7. Do you think, with a proper work from home environment, productivity can improve? - productivity

8. On a scale from 1 to 5 rate your productivity after the work from home culture started- productivityrating
9. How has work from home impacted your stress levels? - stress
10. Has the quality of work decreased because of work from home culture? -workquality
11. Has the salary decreased because of work from home culture? - wfhsalary
12. Has the working hours increased because of work from home culture- wfhtimechange
13. Mention one strong positive point about work from home culture. (single word answer)', wfhpositive Mention one negative point about work from home culture (single word answer)', ' wfhnegative
14. Which age group should be specifically allotted the work from home culture?', ' wfhprefferedage
15. What sort of a work from home job do you prefer?', ' jobpreference
16. Do you have a healthy work-life balance?', w-l_balance
17. Please mention the number of hours you prefer to work for at home. - workhours
18. On a scale from 1 to 5, how well equipped are you to work from home (all factors included from network, power components to all electronic devices and proper workstations)? - equiplevel
19. To promote a healthy and productive work from home culture, what do you think should be the must have work from home essentials? - wfhessentials
20. What is the biggest challenge when it comes to working from home? -wfhchallenge
21. What matters the most, while purchasing a work from home product? - wfhproductfactor
22. Do you agree that the work from home culture can be implemented eventually? -wfhlongrun

2.4 DATA CLEANING

1. The .csv file was obtained from Google Forms directly.
2. Time stamp was removed.
3. The names of all the columns were changed from questions to a column name of choice.
4. Missing/wrongly inputted values were removed.
5. All the extra columns not suitable for modelling and study were removed like rating of the survey and comments on the survey.

6. categorical values have been converted to numeric by either one hot encoding or just their binary forms(eg: yes=1, no=0 or male=1, female=0)

3. ANALYSIS

3.1. EDA

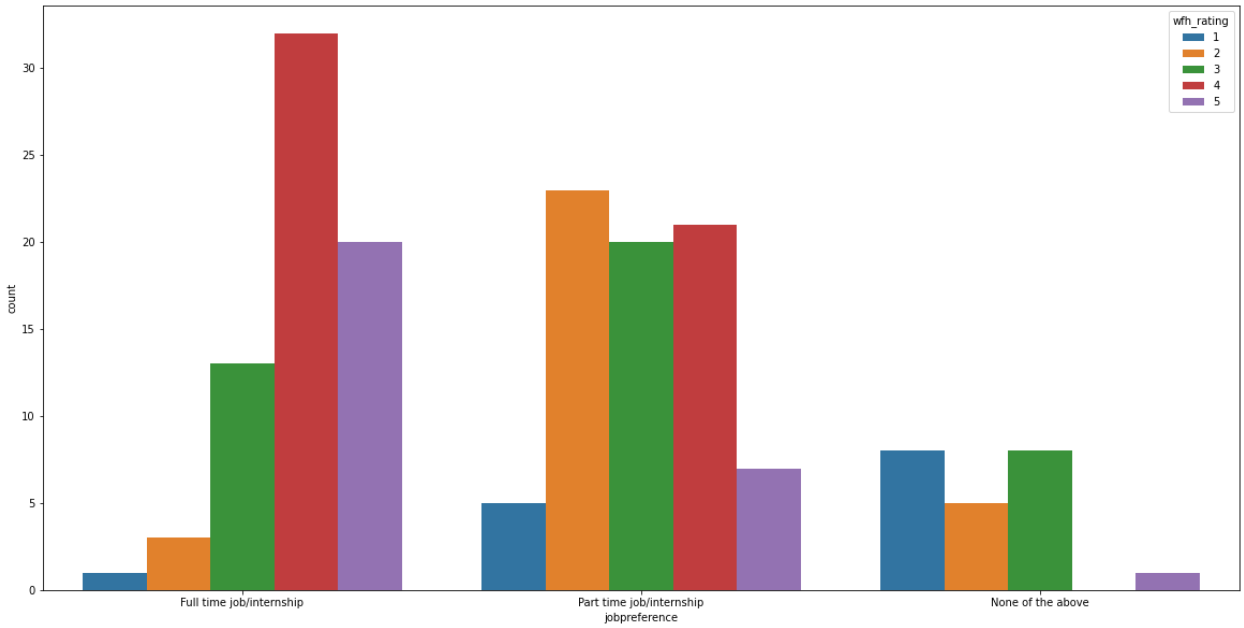

Fig 1: Countplot for job preference Factored by wfh rating

From this graph, we can see that people who prefer wfh at the higher end also prefer more of a full time job/internship.

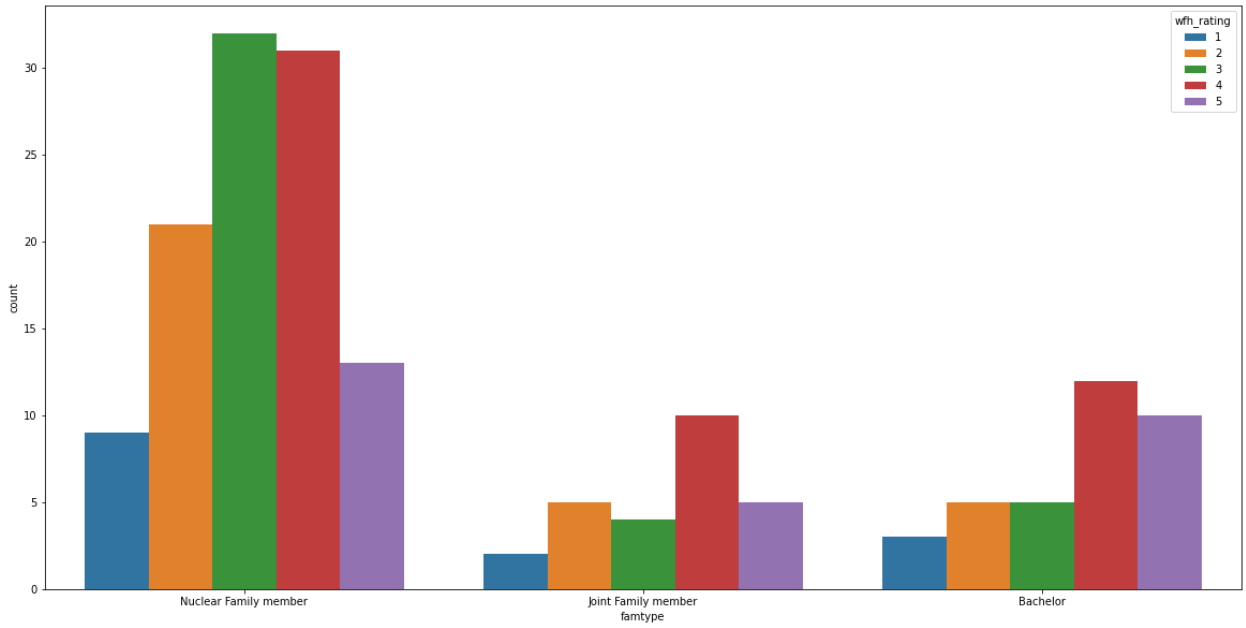

Fig 2: Countplot for famtype Factored by wfh rating

The above graph shows that more of people who are a Nuclear Family member prefer work from home culture.

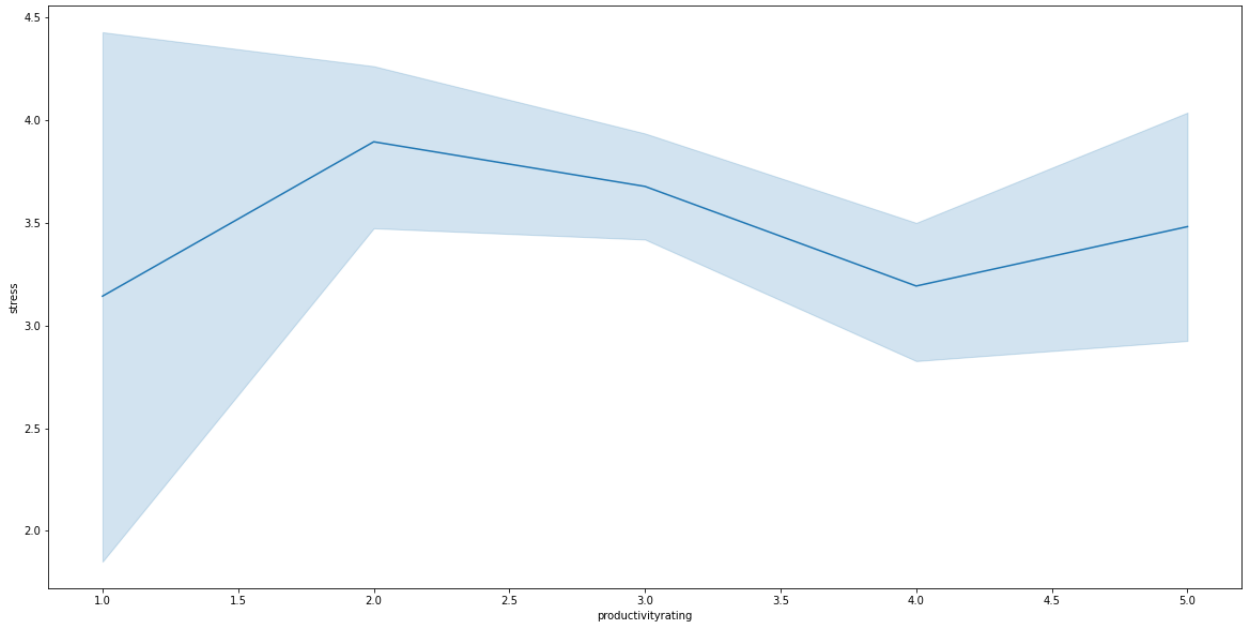

Fig 3: Line Plot for productivity vs stress

The above graph shows a relationship between productivity and stress. Due to less no of responses, we aren't able to confer if it follows a linear or non linear relationship.

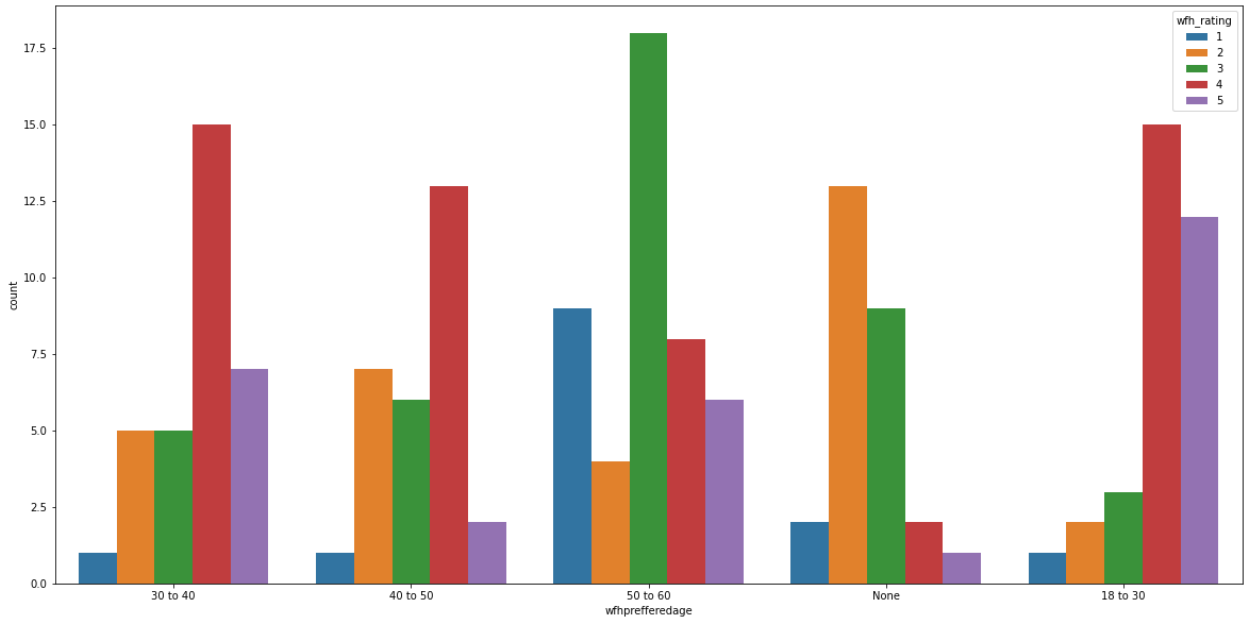

Fig 4: Countplot for work from home preferred age factored by wfh rating

The above plot shows that more people belonging to the age group 50 to 60 should be allowed to work from home compared to the rest of the age groups.

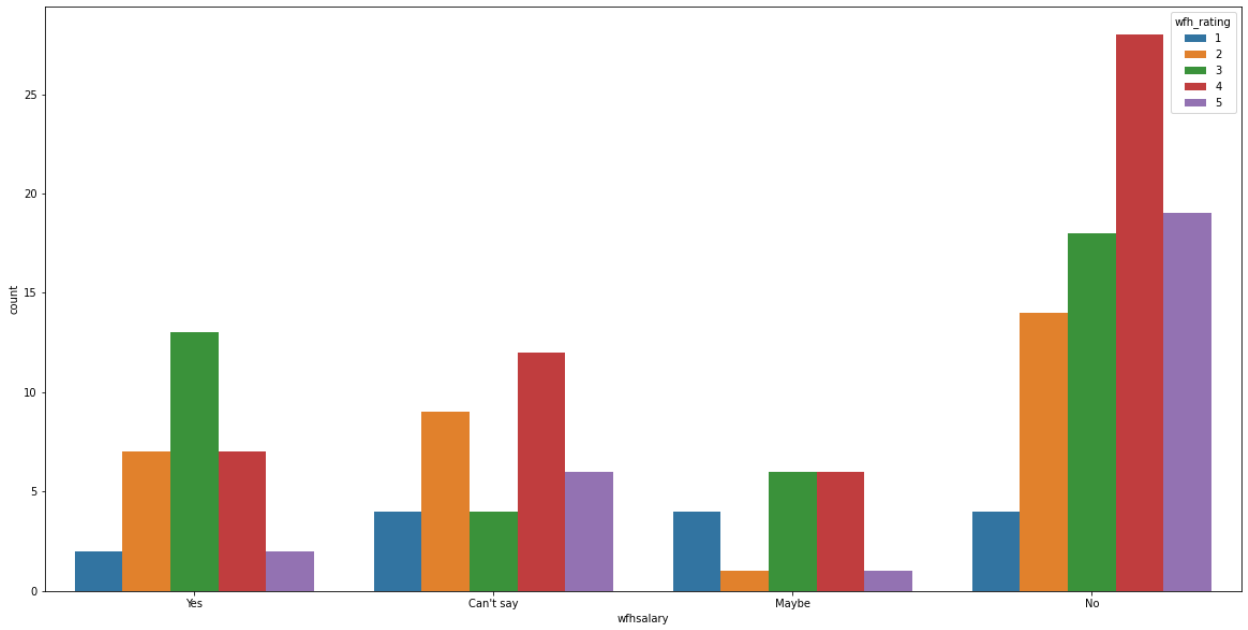

Fig 5: Countplot for wfh salary factored by wfh rating

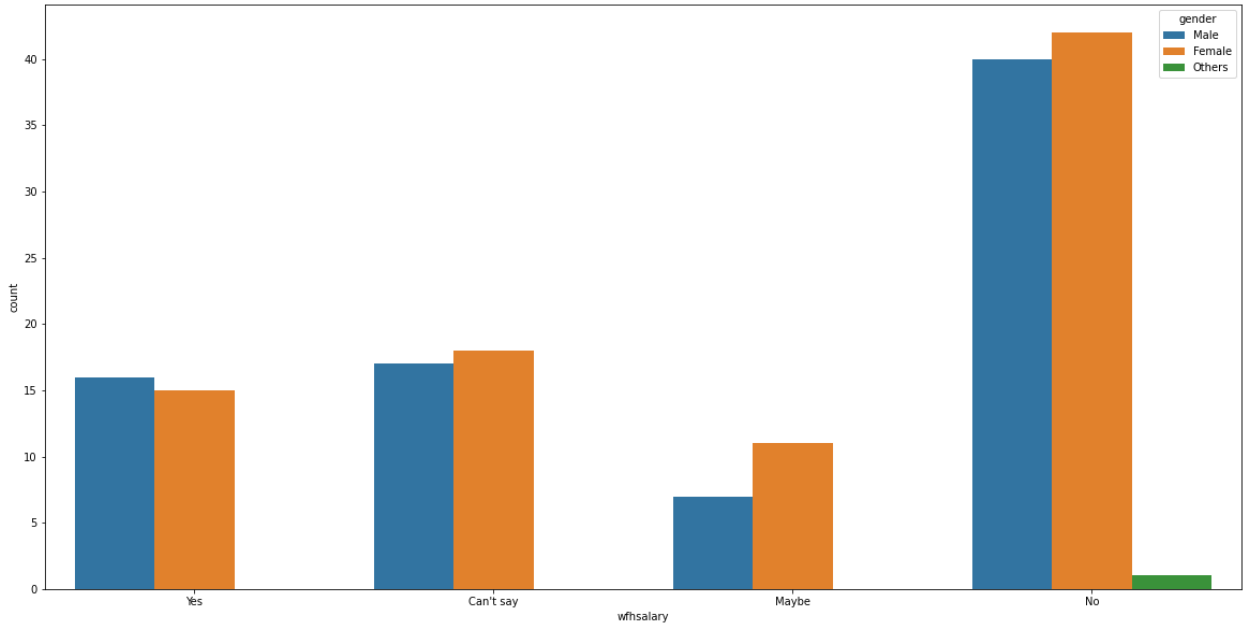

Fig 6: Countplot for wfh salary factored by gender

There was a question in the survey asking if the salary has decreased because of work from home culture. From the plot we can see that, majority of the respondents deny the fact that there has been a decrease in the salary due to work from home culture. From the second plot,(Fig 6), we can see that more no of female respondents have denied the fact.

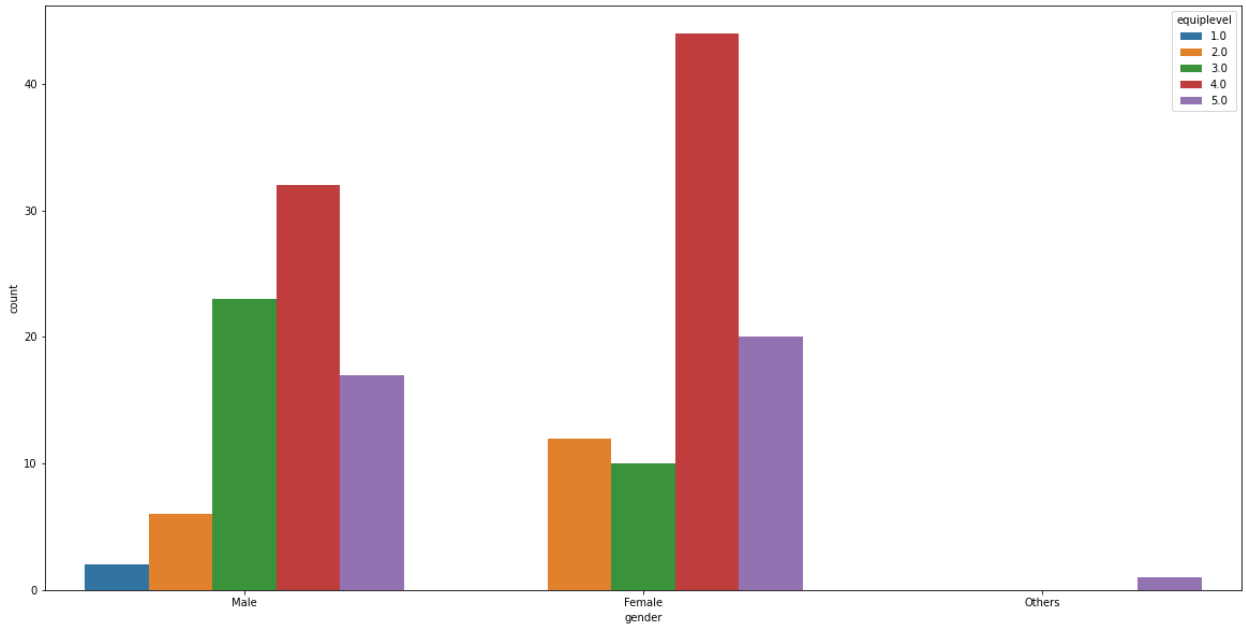

Fig 7: Countplot for gender factored by equiplevel

From this graph, we can see that more no of female respondents are well equipped with work from home products compared to the rest. This also shows that they are more inclined towards working from home.

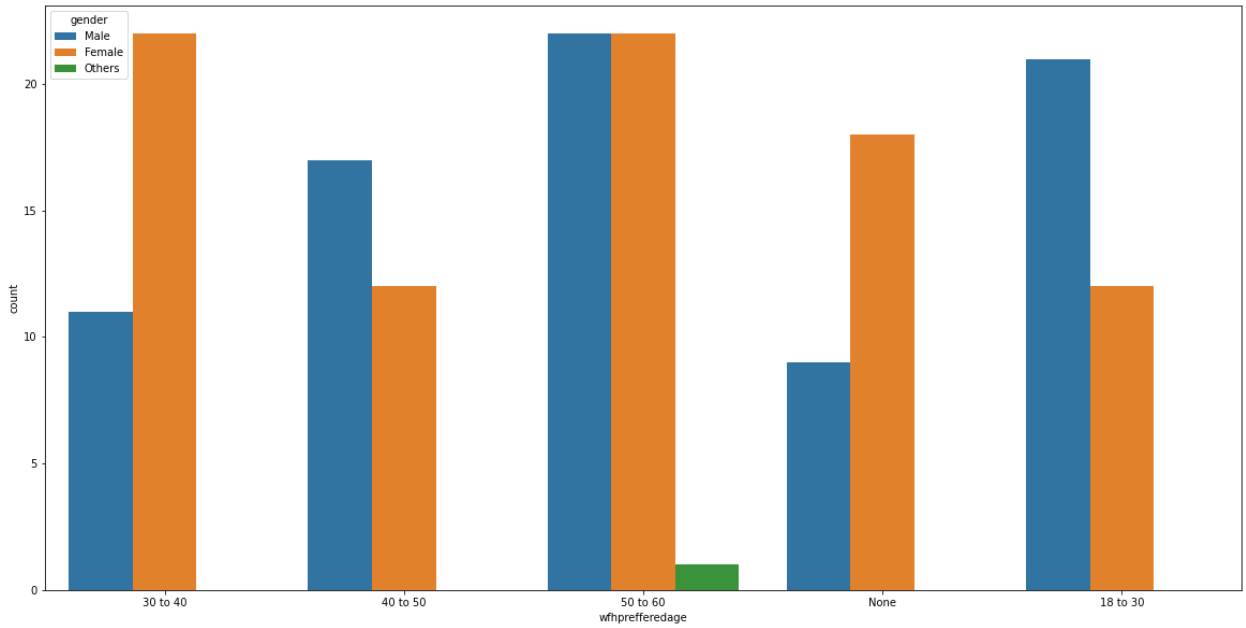

Fig 8: Countplot for wfh preferred age factored by gender

We can see that in the age group 30 to 40, females prefer work from home more than males. In the age group 40 to 50 and 18 to 30, males prefer to work from home more than females. The age group that prefers to work from home the most, i.e., 50 to 60, has an equal ratio for both males and females.

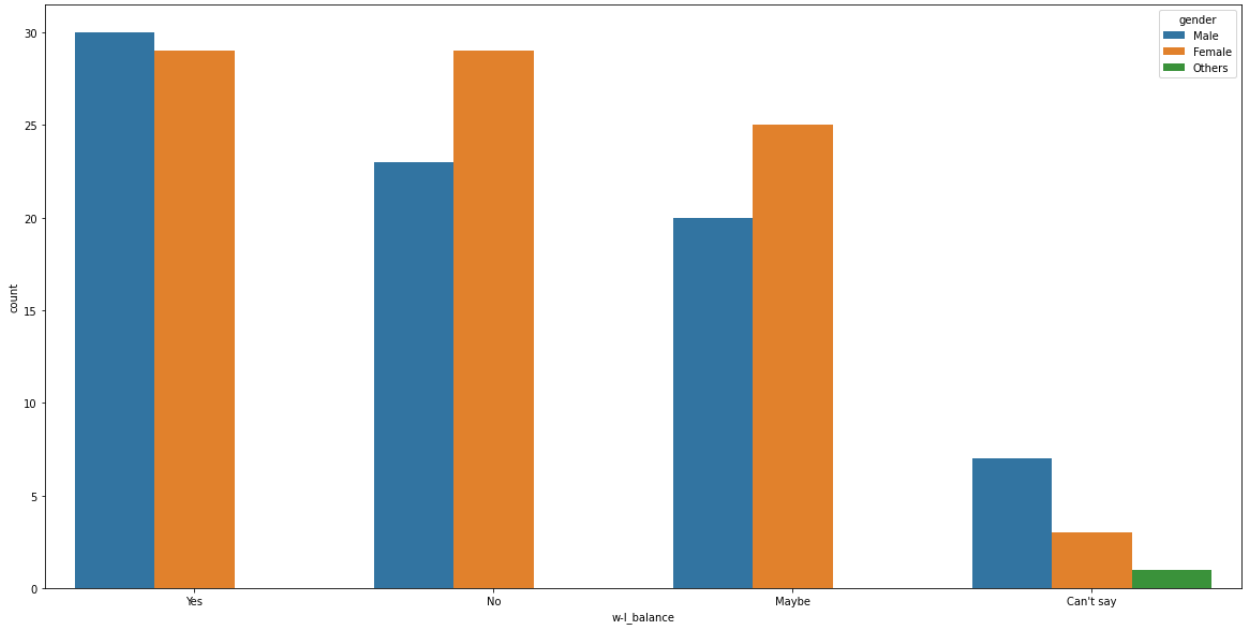

Fig 9: Countplot for work life balance age factored by gender

There was a question in the survey that if due to the work from home culture, respondents are able to have a healthy balance between work and life. We can see that more of the female respondents have denied the fact that there is a work life balance compared to males. But, an equal proportion of female respondents have agreed as well and a higher proportion of males have also agreed to have a proper work life balance.

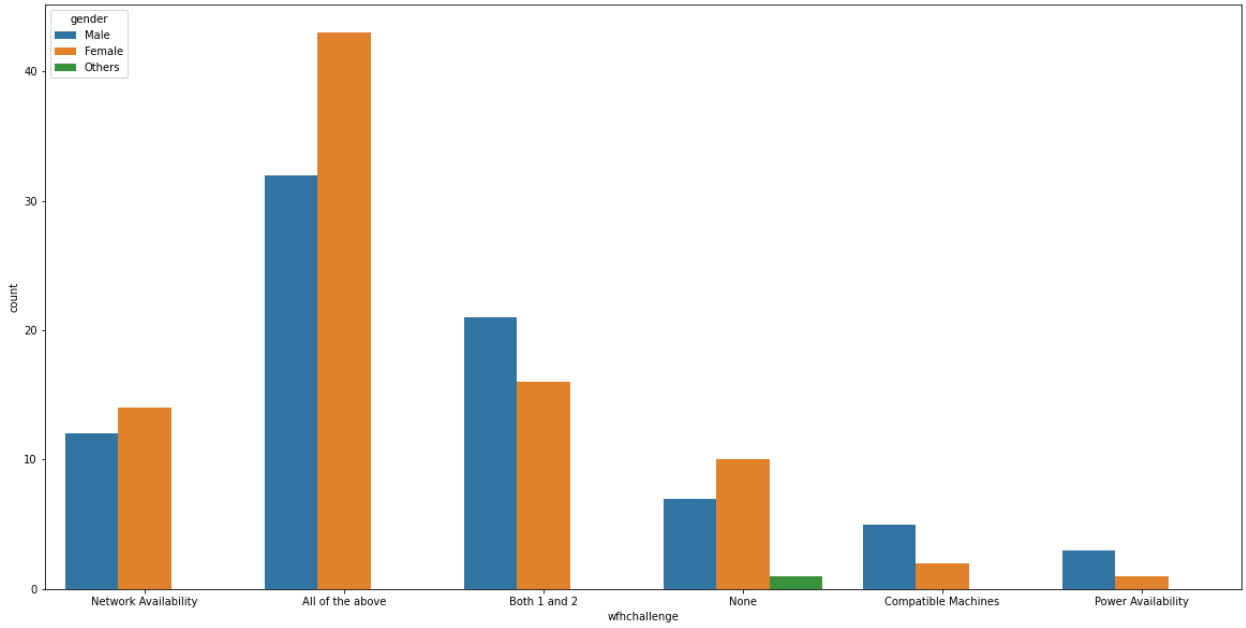

Fig 10: Countplot for work from home challenges factored by gender

We can see that we have a majority of the female respondents who accept that Network Availability, Power Availability and Compatible Machines are all important challenges to solve in a work from home culture.

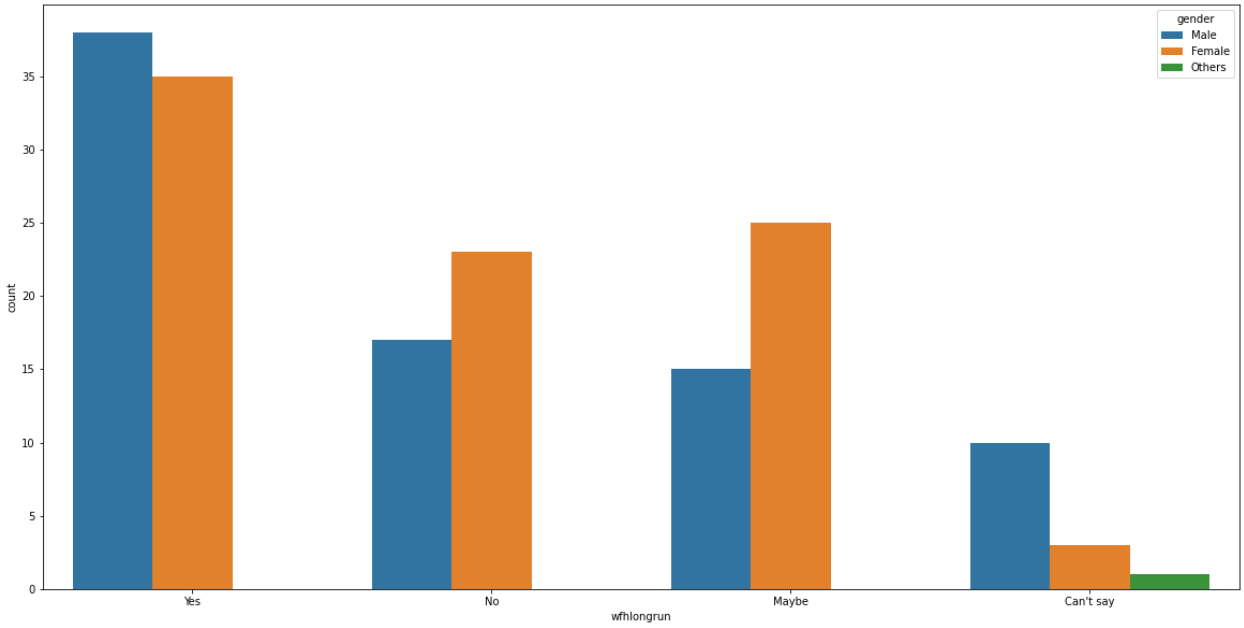

Fig 11: Countplot for wfhlongrun factored by gender

The survey had asked respondents if work from home can be continued in the long run. We can see that a good proportion of both male and female respondents have agreed that work from home can be implemented in the long run.

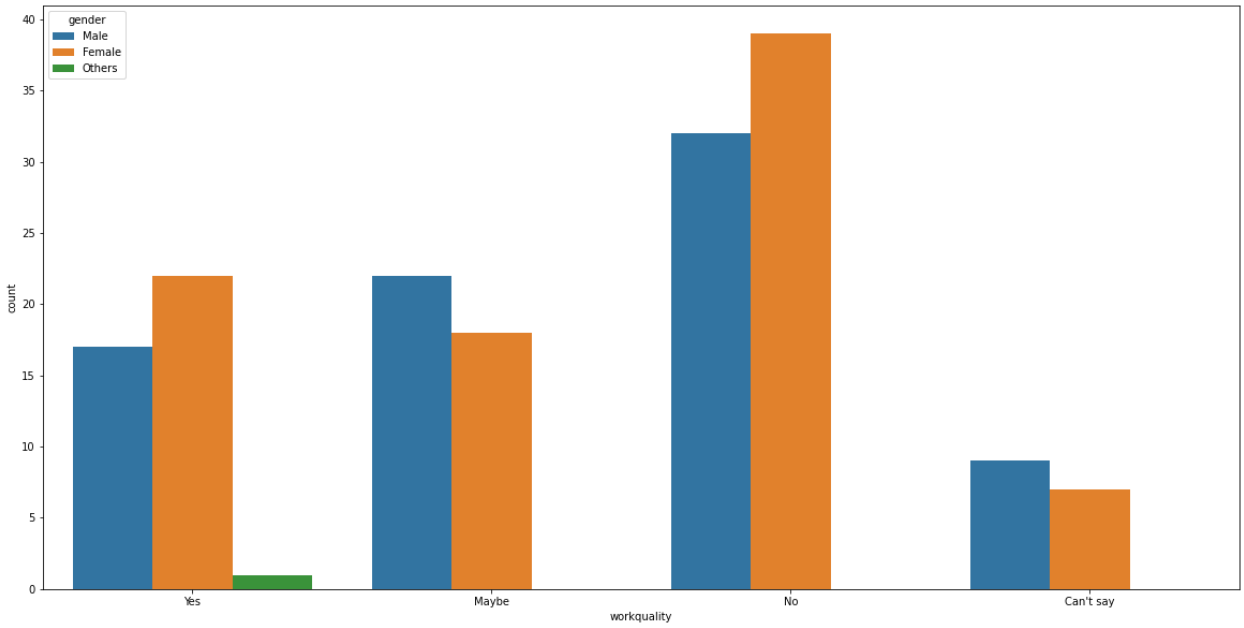

Fig 12: Countplot for work quality factored by gender

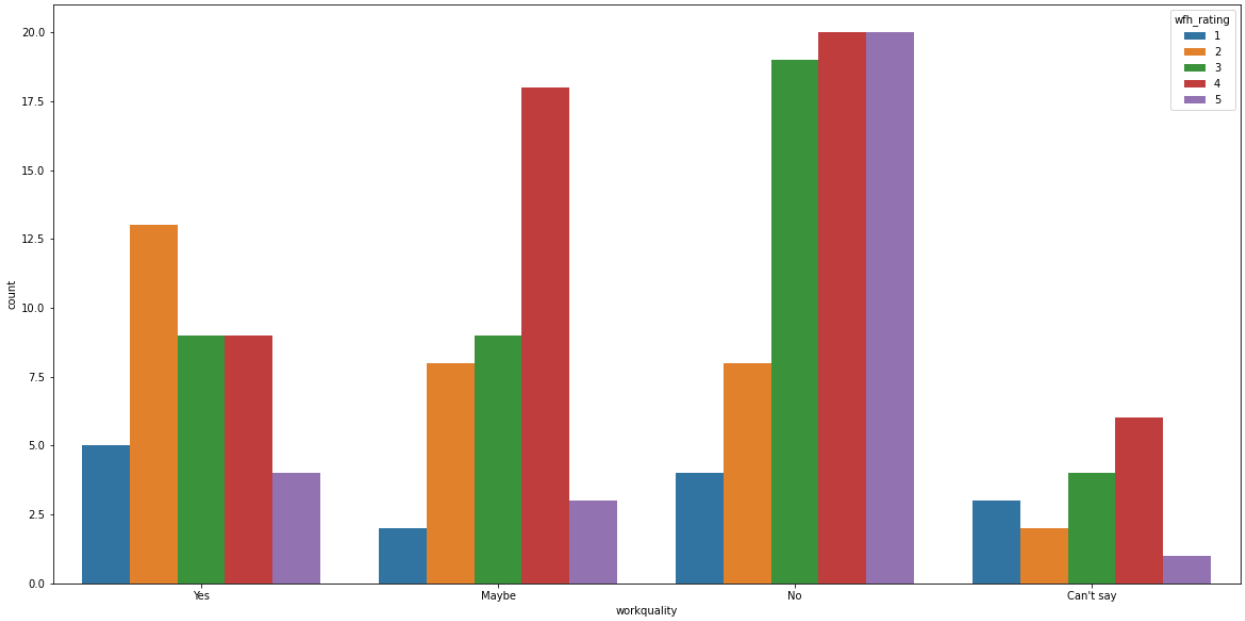

Fig 13: Countplot for work quality factored by work from home rating

The survey had asked the respondents if the work quality has decreased due to the work from home culture. We can see that a good proportion of both male and female respondents, who in turn prefer work from home culture, highly, have denied the fact that work quality has decreased due to the work from home culture.

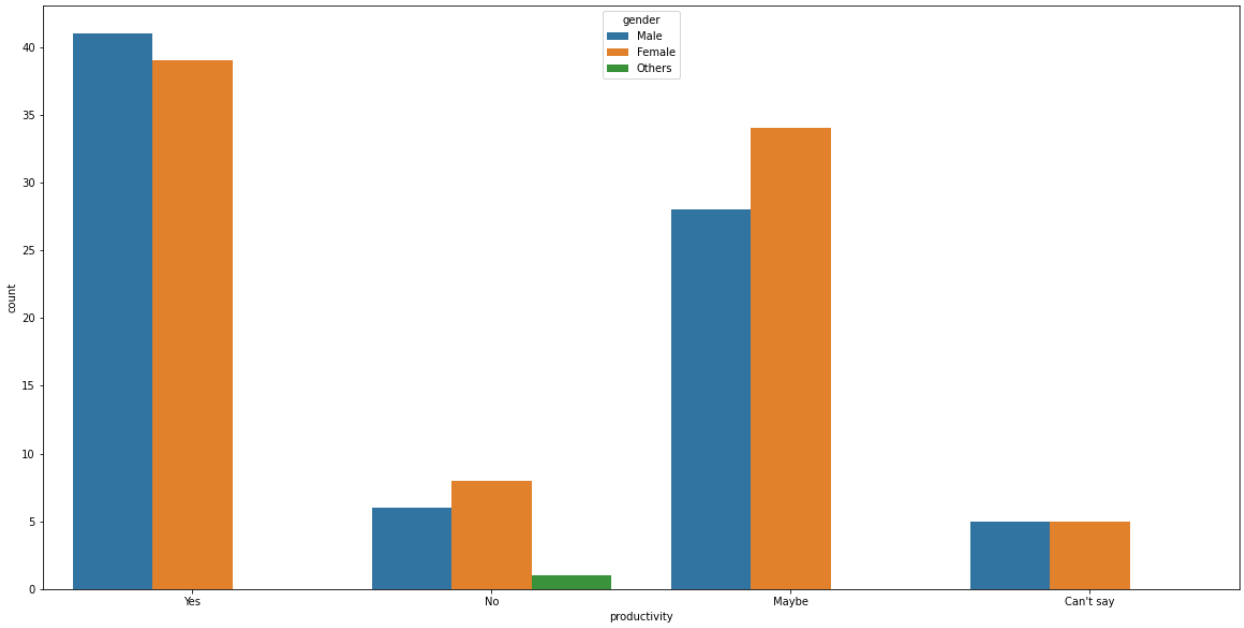

Fig 14: Countplot for productivity factored by gender

The survey had asked the respondents whether a proper work from home culture can bring improvement in productivity. We can see that a good proportion of both male and female respondents agree that a proper work from home culture can improve the productivity levels.

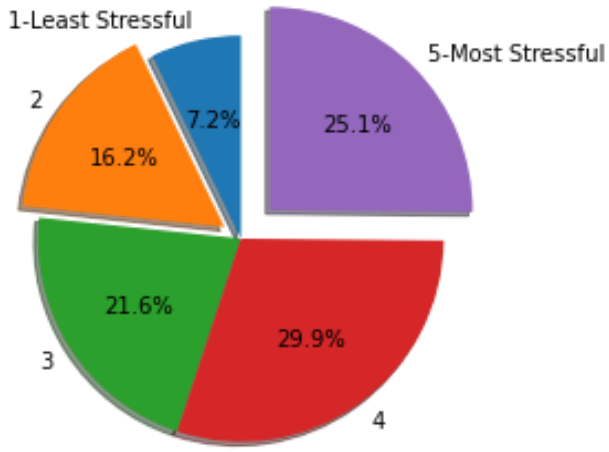

Fig 15: Pie chart showing how stressful work from home culture is

The above graph shows that more no respondents find it stressful to work from home.

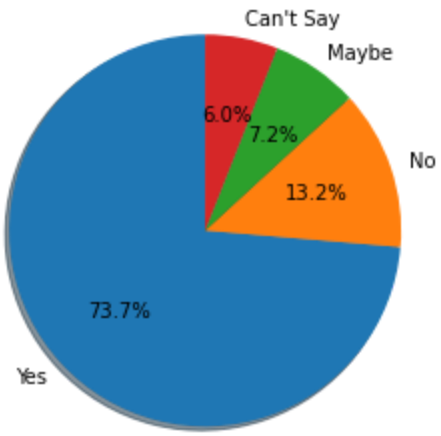

Fig 15: Pie chart showing whether working hours have increased due to work from home.

The above graph shows that a higher percentage of the respondents agree to the fact that the working hours have increased due to work from home culture.

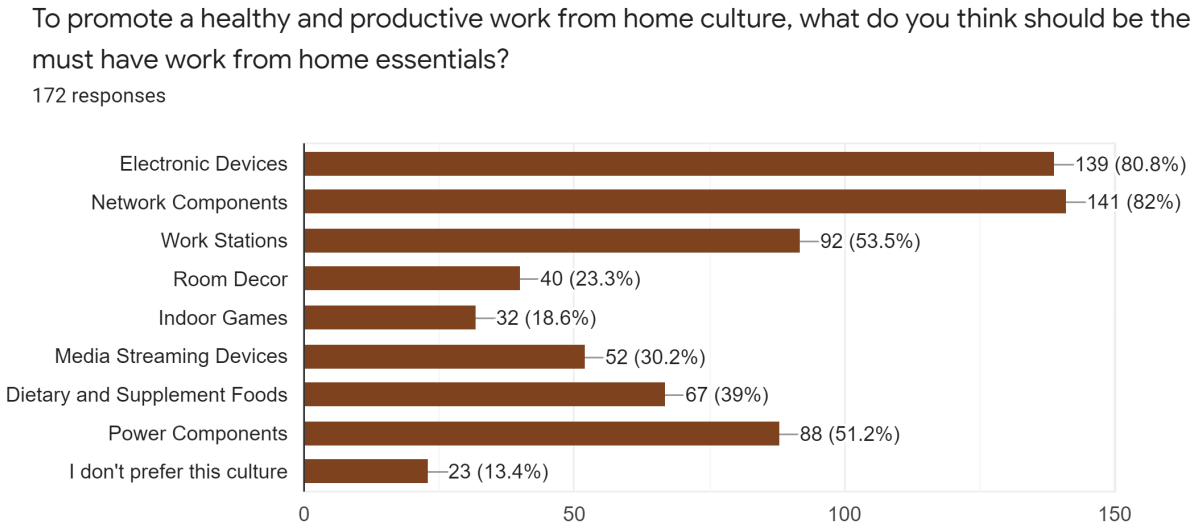

Fig 16: Bar Chart Showing which are the important work from home essentials.

The above chart collected from Google Forms shows that people are more concerned with having proper electronic devices and network components rather than any other factor.

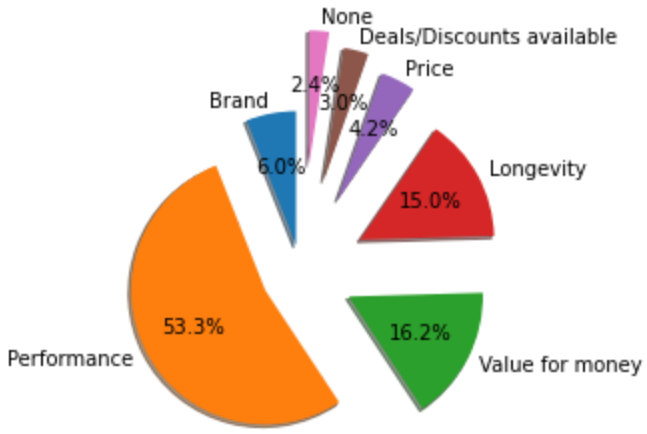

Fig 17: Pie Chart Showing which are the important factors to consider while considering work from home essentials.

The above chart shows that more no of respondents consider Performance as an important factor while looking for a work from home essential product.

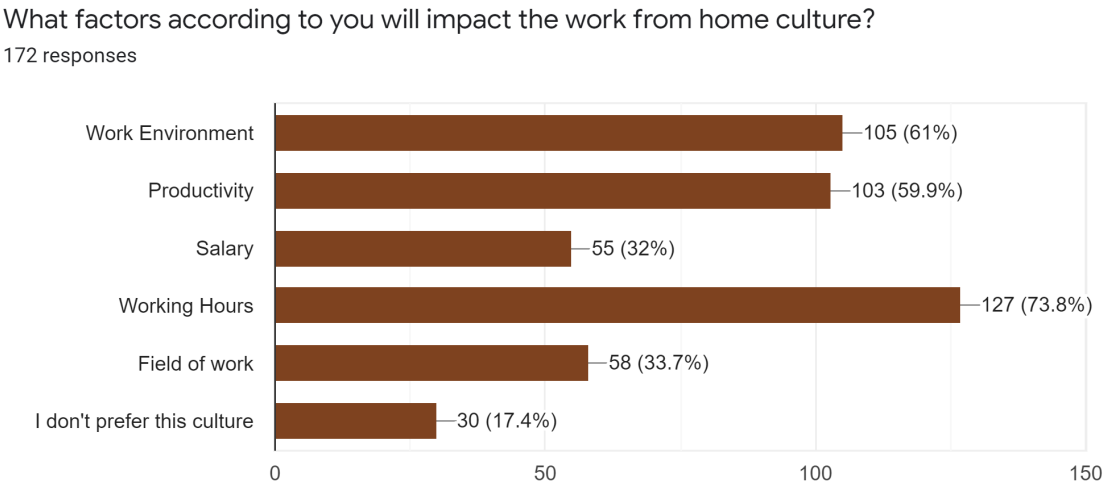

Fig 18: Bar Chart Showing which are the important work from home factors.

The above chart shows that Working hours impact the work from home culture the most along with work environment and productivity.

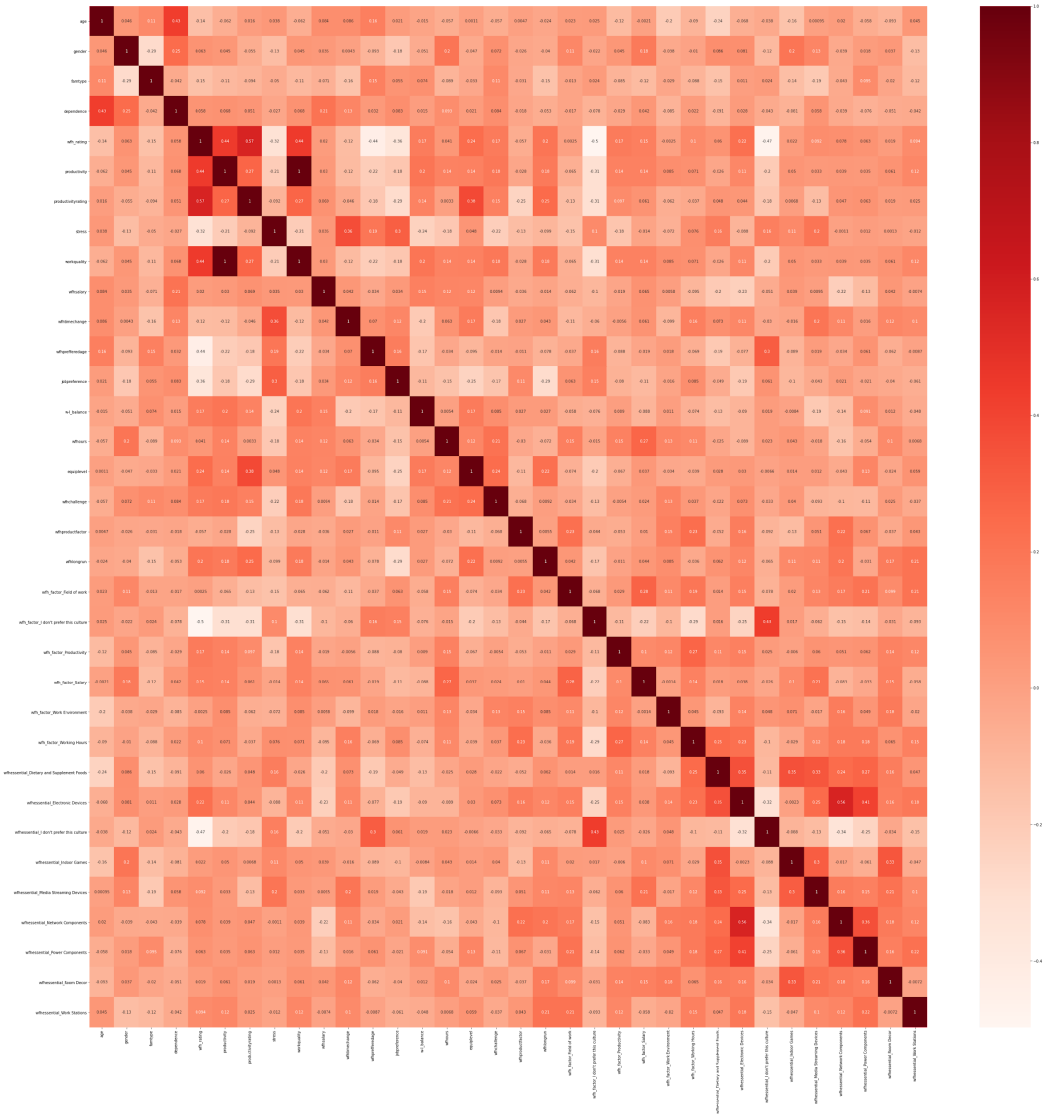

Fig 19: heatmap

From this graph not only do we get to see what has the highest correlation with the dependent variable but also, we get to see that a lot of the independent variables are correlated to each other, and this is called multicollinearity.

Another inference that can be made is that even though there are a lot of independent variables only a few of them truly effect the dependent variable.

3.2 WORD CLOUD ANALYSIS

The first step in our assessment is to see all the positives and negatives of the work from home culture and for this we have used word cloud.

In our survey we made sure to ask about the positives and negatives in 1 to 2 words about this work from home culture, this was done so that we could try and see what is the most common positive and negative aspect of this culture from the public's eye.

1. Positive characteristics about work from home culture:

```
comment_words = ''
stopwords = set(STOPWORDS)

# iterate through the csv file
for val in df.wfhpositive:

    # typecaste each val to string
    val = str(val)

    # split the value
    tokens = val.split()

    # Converts each token into lowercase
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()

    comment_words += " ".join(tokens)+" "

wordcloud = WordCloud(width = 1000, height = 1000,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (20, 10), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

Fig 20: Positive comments about work from home

We can see those words like flexibility, comfort, time, savings, family have been the ones that have been used the most, this means that the these are the positives offered by the work from home environment.

2. Negative characteristics of working from home:

Fig 21: Negative Comment about work from home.

Here we can see stress, time, boring, unproductive, less and work being the most used words this would mean that these are the negative characteristics.

So even though work from home provides us with flexibility, from this analysis we can note that, they must work for longer, and more than they used to. There are perks when it comes to working from home but at the same time there are a lot of cons as well.

4. MODEL CONSTRUCTION

We can construct models in 2 ways (using Multinomial logistic Regression model):

1. Keeping all the features and using algorithms
2. Using algorithms after feature selection

1. Keeping all the features and using algorithms

We used Multinomial logistic regression technique without feature selection and this was the classification report

```
print(classification_report(yhat1,y_test))
```

	precision	recall	f1-score	support
1	0.25	0.33	0.29	3
2	0.56	0.62	0.59	8
3	0.36	0.36	0.36	14
4	0.69	0.58	0.63	19
5	0.62	0.71	0.67	7
accuracy			0.53	51
macro avg	0.50	0.52	0.51	51
weighted avg	0.54	0.53	0.53	51

Fig 22: Classification report for Multinomial Logistic Regression 1

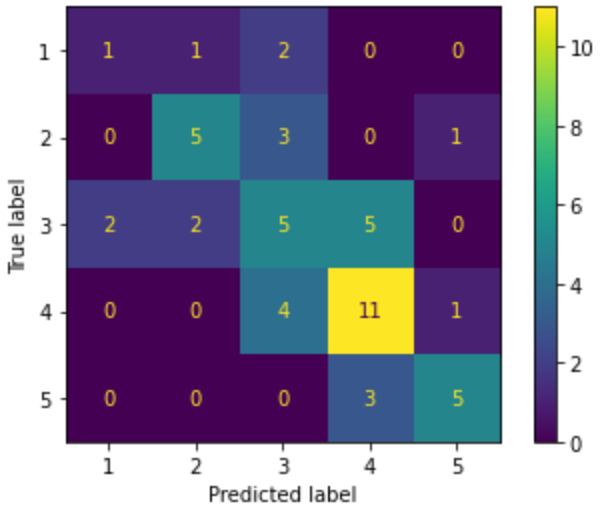

Fig 23: Confusion matrix for Multinomial Logistic Regression 1

We can see that we have achieved an accuracy of 53 percent which can be improved.

2. Now let us try with feature selection technique

Backward Feature Elimination: This method works exactly opposite to the Forward Feature Selection method. Here, we start with all the features available and build a model. Next, we the variable from the model which gives the best evaluation measure value. This process is continued until the preset criterion is achieved.

List of selected variables:

['age',
'famtype',
'dependence',
'productivity',
'productivityrating',
'wfhsalary',
'wfhtimechange',
'wfhprefferedage',
'jobpreference',
'w-l_balance',

'wfhchallenge',

"wfh_factor_I don't prefer this culture",

'wfh_factor_Productivity',

'wfh_factor_Work Environment',

'wfh_factor_Working Hours',

"wfhessential_I don't prefer this culture",

'wfhessential_Media Streaming Devices',

'wfhessential_Power Components',

'wfhessential_Room Decor']

The following is the classification report

```
print(classification_report(yhat2,y_test1))
```

	precision	recall	f1-score	support
1	0.00	0.00	0.00	1
2	0.89	0.80	0.84	10
3	0.57	0.53	0.55	15
4	0.75	0.67	0.71	18
5	0.62	0.71	0.67	7
accuracy			0.65	51
macro avg	0.57	0.54	0.55	51
weighted avg	0.69	0.65	0.67	51

Fig 24: Confusion matrix for Multinomial Logistic Regression 2

We can see that we have achieved a greater accuracy with Backward Feature Selection Technique with an accuracy of 65 percent. Since we have lesser entries of data, the accuracy is low and can be surely increased with more data.

CONCLUSION

A model will always improve after dimensionality reduction.

The most prominent features in this model are:

['age',
'famtype',
'dependence',
'productivity',
'productivityrating',
'wfhsalary',
'wfhtimechange',
'wfhprefferedage',
'jobpreference',
'w-l_balance',
'wfhchallenge',
"wfh_factor_I don't prefer this culture",
'wfh_factor_Productivity',
'wfh_factor_Work Environment',
'wfh_factor_Working Hours',
"wfhessential_I don't prefer this culture",
'wfhessential_Media Streaming Devices',
'wfhessential_Power Components',
'wfhessential_Room Decor']

The following is the classification report

From the above we can note that:

Productivity is the most important factor.

The most essential requirements are:

1. electronic devices
2. indoor games
3. media streaming devices
4. network
5. electricity
6. decor
7. work station.

In a case like ours where there exists a lot of multicollinearity and very few crucial features, lasso regression works very well.

From the word cloud analysis, we got to note what is the best part about work from home and what is the worst part about the same, which helps us to make work from home better.

We can safely conclude that Multinomial Logistic Regression with Backward Feature Selection technique gave us a better score which can be improved with more no of entries.

LIMITATION OF STUDY

1. Size: The size of the data set was just 166 rows, which is exceedingly small and that was definitely a limitation.
2. Time: If we had more time to work on the project, we could have given better results.
3. Human error: While people filled in the survey, they made a couple of errors here and there.

REFERENCES

1. **https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/**
2. **https://machinelearningmastery.com/multinomial-logistic-regression-with-python/**

ACKNOWLEDGEMENT

The authors of this article would like to express their sincere gratitude to our mentor and guide Dr. Ummesalma M for providing the opportunity to carry out this research assignment as part of the CIA work for MDS-231 Machine learning. We thank you ma'am for the constant support and encouragement. We also like to thank all the respondents who have taken their time to respond to the questionnaires being sent and helping the authors to produce new insights and results on the objectives considered. We would also like to express our sincere gratitude to our parents for their constant support during the process of this project. We would also like to thank our friends without which this would not have been possible. Last but not the least we thank and invoke the blessings of the almighty.