**AS1202 Advanced Statistics**

**Analysis of People’s choice on working mode**

PREPARED BY

**Siddharth Jangid (2020BTechCSE074)**

FACULTY GUIDE

**Dr. Jaya Gupta**

**Logo, company name

Description automatically generated**

**Department of Computer Science and Engineering**

**Institute of Engineering and Technology**

**JK Lakshmipat University Jaipur**

**December 2022**

# CERTIFICATE

This is to certify that the project work entitled “**Analysis of People’s choice on work mode**” submitted by **Siddharth Jangid** towards the partial fulfilment of the requirements for the degree of **Bachelor of Technology in Computer Science Engineering** of JK Lakshmipat University, Jaipur is the record of work carried out by them under our supervision and guidance. In our opinion, the submitted work has reached the level required for being accepted for the final submission of the project.

|  |
| --- |
|  |
| Dr. Jaya Gupta |
| Assistant Professor-Mathematics |
| Institute of Engineering & Technology  JK Lakshmipat University, Jaipur |
|  |
|  |

Date of Submission: December 2022

# ACKNOWLEDGEMENTS

I have completed this project under the guidance and supervision of **Dr. Jaya Gupta** Assistant Professor, IET, JK Lakshmipat University, I will be failed in my duty if I do not acknowledge the esteemed scholarly guidance, assistance, feedback, and knowledge I have received from them towards the fruitful and timely completion of this work.

I express our deepest thanks to **Dr Dheeraj Sanghi**, Vice Chancellor, JK Lakshmipat University, and **Dr Sanjay Goel**, Director, Institute of Engineering and Technology, JK Lakshmipat University for their constant support, encouragement, and guidance.

I also acknowledge with a deep sense of reverence, our gratitude towards our parents for their direct or indirect support during the entire course of this project.

Thanking You

Sincerely Yours,

Siddharth Jangid

# PROJECT SUMMARY

The purpose of this project is to investigate the factors that may affect an employee’s preference for working from home (WFH) or working from the office (WFO). This project will involve collecting data from a survey of multiple people, where the survey responses will be used to generate a dataset that includes features such as age, job type, kids and location etc. This dataset will then be used to create a model that will predict an individual’s preference for WFH or WFO based on the features in the dataset. Participants in the survey will be asked to answer questions related to their job type, age, location and any other relevant demographic or lifestyle details. The survey responses will then be used to generate a dataset which will include the following features: age, job type, salary, location, and other demographic information. The survey responses will also be used to determine the target variable, which will be an individual’s preference for WFH or WFO. Once we have our dataset, we will be able to use it to build a model that can predict an individual’s preference for WFH or WFO. I will be using a variety of machine learning algorithms to construct the model, such as logistic regression, linear regression, one way anowa, and other supervised learning algorithms. Once the model is built, it will then be tested to see how well its predictions match up with the actual results of the survey.

This project is a great opportunity to explore how different factors may play a role in an individual’s preference for working from home or going to the office. It can also be used to gain a better understanding of how an individual’s job type, salary, age, location and other demographic details can influence their decision to work from home or the office. By understanding the factors that can impact an individual’s decision, employers and managers can better design policies that take into account the needs of their employees.

**CONTENTS**

**Page No.**

**CHAPTER 1: GOAL AND MOTIVATION 6.**

**CHAPTER 2: UNIQUENESS AND USP OF YOUR PROJECT 7.**

**CHAPTER 3: OBJECTIVE 8.**

**CHAPTER 4: METHODOLOGY 9.**

**CHAPTER 5: RESULT AND DISCUSSION 14.**

**CHAPTER 6: CONCLUSSION 26.**

**REFERENCES 27.**

# CHAPTER 1: GOAL AND MOTIVATION

The goal of this project is to explore the impact of working from home (WFH) and working from office (WFO) on the satisfaction of individuals within organizations. Working from home has become increasingly popular in recent years due to the advancements in technology and internet accessibility, with many organizations allowing employees to work remotely. While this certainly provides unprecedented flexibility, there are still concerns regarding the impact such an arrangement can have on an individual's efficiency and overall satisfaction.

The motivation for this project is to better understand how working from home and working from office affects the satisfaction levels of individuals. This information can be used to make sound decisions within an organization and understand how the needs of their employees can be met while still promoting productive and efficient places of work. The survey taken and data collected are the foundation of this project. Multiple people were asked a series of questions in order to assess their preferences when it comes to WFH and WFO, and their answers will be used as the basis for the project. Regression and one-way ANOVA analysis will be used to compare the actual and predicted values from the data to further analyze the satisfaction with WFH and WFO working arrangements.

The significance of this project lies in being able to assess which working arrangements provide the greatest satisfaction for individuals. Being able to answer this question makes it easier for organizations to understand the needs of their employees and make the necessary adjustments in order to foster an environment conducive to productivity. In addition, this project has the capability of creating a benchmark for which other organizations can measure the effectiveness of their working arrangements and create strategies that further ensure employees' satisfaction. In conclusion, this project is a crucial step in understanding how working from home and working from office has an effect on the satisfaction of individuals in an organization. The collected survey data is the starting point for building a successful foundation that can be used to answer the question of which working arrangement yields the highest satisfaction.

# CHAPTER 2: UNIQUENESS AND USP OF YOUR PROJECT

This project seeks to use linear regression and one-way analysis of variance (ANOVA) analysis to determine the efficacy of predicting the preference of working from home or working from office (WFH vs WFO) in a sample of survey respondents. The project has identified a specific set of survey questions and variables that can be used to classify the respondents' preferences, offering an objective and actionable way to predict their outcome.

The use of linear regression and ANOVA will help to identify any patterns or relationships in the data that may otherwise remain undetected. Furthermore, the insights gained from this project will provide a better understanding of the factors that influence a person's preference for working from home or office. This knowledge can help to improve the experience of both individuals working from home and the organizations offering the opportunity.

The results of the project can potentially inform key decisions related to the recruitment, retention, and work-life balance of employees. Ultimately, this research project provides a unique opportunity to gain valuable insights from an empirical data source, which can ultimately empower both employers and employees in making smarter decisions.

# CHAPTER 3: Objective

1. To analyze dataset using Python programming language.
2. Statistical analysis

* To analyze what people favors between work from home and work from office on the basis of Gender.
* To analyze what people favors between work from home and work from office on the basis of productivity of individual.
* To analyze what people favors between work from home and work from office on the basis of office location.
* To analyze what people favors between work from home and work from office on the basis of employee who have kids.
* To analyze what people favors between work from home and work from office on the basis of Age Group.
* To analyze what people favors between work from home and work from office on the basis of state of mind(calmer/stressed).

1. To calculate correlation coefficient of each column with another column of dataset.
2. To predict accuracy of Target value of random employee from dataset using logistic regression and confusion matrix.
3. To predict who will have a better work life balance on the basis of different factors by using multiple linear regression.
4. To determine the existence of a statistically significant difference among several group means by using One-way anova.
5. To estimate how the mean of a quantitative variable changes according to the levels of categorical variables(RM\_productive).

# CHAPTER 4: Methodology

The methodology of the project above involves a survey and data analysis to determine whether people prefer working from home or working from an office. The survey consisted of multiple questions and was used to collect data from a variety of people across different age groups, genders, and job titles. To begin the process, the survey was distributed to a sample of participants across the different demographic groups. This sample was selected to adequately represent the entire target population to ensure that the data gathered would be a reliable measure of worker preferences.

Once the responses were collected, the data was entered into a database and analyzed using a variety of statistical methods, including linear regression, one way ANOVA, two way ANOVA, and correlation. These analytical tools allowed for the correlation of data between questions and responses, as well as for comparison between different demographic groups. Next, the analytical results were interpreted to make predictions about which workers prefer to work from home or from an office. The results of the analysis showed that age, job title, and gender had a significant effect on the likelihood of workers’ preference for working from home or from an office. Finally, to confirm the accuracy of the predictions, the results of the analysis were compared to the target responses that the survey participants had originally provided. By measuring the correlation between the results of the analysis and the target responses, it was possible to verify that the predictions were correct.

Overall, the methodology employed in this project allowed for the collection of reliable data and the analysis of that data to make accurate predictions about people’s preferences for working from home or from an office. This knowledge can be helpful in making decisions about how to best utilize office workspace as more people explore the possibility of telecommuting and working from home.

**Correlation:** Correlation can be defined is an analysis of co-relation between two or more variable such as x & y.

Types of correlation:

1. Positive correlation: positive correlation comes into picture when both the variable deviated in same direction. i.e. The variable moves in same direction
2. Negative correlation: Negative correlation comes into picture when both the variable does not deviate in same direction. i.e. The variable does not move in same direction.
3. No correlation: No relation in the variable x & y
4. Linear correlation: The relation of variable in which x & y grow linearly i.e. y= mx + c
5. Non-linear correlation: The relation of variable in which x & y does not grow linearly. i.e. y= x2 + 1.
6. Graphical user interface, text, application

   Description automatically generatedWhere r is coefficient of correlation.

Properties of co-relation:

* If r > 0 then it is +ve correlation
* If r >0 then it is -ve correlation
* If r = 0 then it is No correlation

Note: Correlation coefficient is independent of change of origin and change of scale.

**Simple linear Regression:** The simple meaning of regression is stepping back or returning to the average value. Regression analysis implies the prediction or estimated of dependent variable if the independent variable value is known.  
 It is a very important statistic tool which is expressively used in all branches of science & management.

Graphical user interface

Description automatically generated

# Multiple linear regression:

# Multiple linear regression is a technique in statistics which uses more than one explanatory variable to predict outcome of a response variable. Multiple linear regression gave us a linear relationship between independent variable and dependent variable. It is an extension of simple linear regression which uses only one independent variable whereas multiple regression can have more than independent variable. This technique mostly used in econometrics and financial inference.

# 

# One-way anova

# One-way Analysis of Variance (ANOVA) is a statistical method used to determine the difference between two or more population means. This technique tests the null hypothesis that the population means are all equal, or that there are no differences between the groups. ANOVA is widely used for testing hypotheses when there are more than two categories of dependent variables or samples.

# The mathematical formula for ANOVA is as follows:

# 

# where μi represents the population mean of the ith group, and k represents the total number of groups.

# The One-way ANOVA is used to understand the differences between means and various sample sizes in order to make inferences about the population means. The goal of the ANOVA test is to determine if there are significant differences between the means of two or more independent groups. To analyze One-way ANOVA, first, a researcher must design the experiment. This consists of arranging the independent variables (IVs) into groups within the parameters of the experiment. Then, the randomization process of allocating test subjects to their assigned IV group must be conducted. Next, descriptive statistics need to be calculated to provide a means of comparison between the independent groups. The mean, median, mode, standard deviation, and other appropriate tests should be calculated and reported. The number of samples within each IV group should also be noted. Then, to determine if there is a significant difference between the means, a hypothesis test should be conducted using the null and alternative hypotheses stated above. A one-way ANOVA can be run in Excel or using other statistical software. Now, to interpret the results of the One-way ANOVA, there are two tables that need to be read. The “ANOVA” table provides the test statistic, degrees of freedom (df) between the IV and the error, and the p-value. A second table, the “regression” table, provides the estimated means for each IV group. If all the groups have an equal mean, the value for the error term should be near zero. Finally, the researcher should determine if there is a significant difference between the means of the groups by comparing the p-value (probability) obtained from the ANOVA to the pre-determined alpha level. If the obtained p-value is less than the alpha level, this indicates that there is a statistically significant difference between the means of the IV groups and that the hypothesis should be rejected.

# In summary, the One-way ANOVA is a powerful tool for determining significant differences between the means of two or more independent variables and their sample sizes. The testing and interpretation of the results can help researchers make informed decisions about the populations that are being studied and also help them draw conclusions about the populations which can then be used to inform scientific and social policies.

# Two Way Anova:

# Two-way ANOVA is a well-known technique used in statistical analysis to compare the means between two different groups of individuals or observations. It is used to discover whether or not there is a statistically significant difference between the two groups on the dependent variable while also taking into account the effects of one or more independent variables. Two-way ANOVA is a type of between-subjects design in that it analyzes the differences between two or more groups of observations. Two-way ANOVA is a statistical method used to compare the means of two or more independent samples of data. The two-way ANOVA is also referred to as a factorial ANOVA as it can be used to test the effect of factors on a response variable. The most commonly used type of ANOVA is a one-way ANOVA which only tests for one factor. It is important to note that the two-way ANOVA is not used to compare the means between groups like a one-way ANOVA but instead can be used to explore the possible interactions between different factors.

# One benefit of the two-way ANOVA that is very important is that it allows us to simultaneously compare the means between two groups on the same dependent variable while taking into account the effect of two or more independent variables. The first step of a two-way ANOVA as with any ANOVA is to first determine if there is a statistically significant difference in the mean between the two groups on the same dependent variable. The null hypothesis states that there is no difference in the means between the two groups, while the alternative hypothesis is that the means between the two groups are statistically significant. The ANOVA test is a statistical method used to determine the validity of the null hypothesis. The way the test works is that if the p-value is less than alpha (alpha usually is taken to be 0.05) then the null hypothesis is rejected and the alternative is accepted stating there is a statistically significant difference in the means between the two groups. If, however, the p-value is greater than or equal to alpha then the null hypothesis is accepted as there is no statistically significant difference in the means between the two groups. If, however, the p-value is less than alpha then the next step is to investigate the source of the difference by looking at the individual main effects and interactions (if any) of the two independent variables. The main effects are the contribution of the two independent variables to the difference in the means between the two groups, and the interaction effect is the potential relationship between the two independent variables. If it is found that the main effects or interactions are the main sources of the difference in the means then this would suggest that the two independent variables are the primary causes or contributors to the difference in the dependent variable between the two groups.

# In summary, two-way ANOVA is a powerful statistical technique used to compare the means of two or more independent groups while taking into account the possible effects of two or more independent variables. It is important to remember that the two-way ANOVA should not be used to confidentially compare the means of two groups, but instead to explore the possible interactions between different factors and their effect on the dependent variable.

# CHAPTER 5: RESULT AND DISCUSSION

# Objective 1:

1. **To analyze dataset using Python programming language.**

# 

# 

# Objective 2 : statistical analysis

1. **To analyze what people favors between work from home and work from office on the basis of Gender**.

# Graphical user interface, text, application, email Description automatically generated

# Chart Description automatically generated

# By the above diagram we can conclude that Womens are more in favor of Work from Office mode.

1. **To analyze what people favors between work from home and work from office based on productivity of individual.**

# Chart, bar chart Description automatically generated

1. **To analyze what people favors between work from home and work from office on the basis of office location.**

# Graphical user interface, text, application, email Description automatically generated

# Chart Description automatically generated

1. **To analyze what people favors between work from home and work from office on the basis of employee who have kids.**

**Graphical user interface, text, application, email

Description automatically generated**

# Chart Description automatically generated

1. **To analyze what people favors between work from home and work from office on the basis of Age Group.**

Table

Description automatically generated

# 

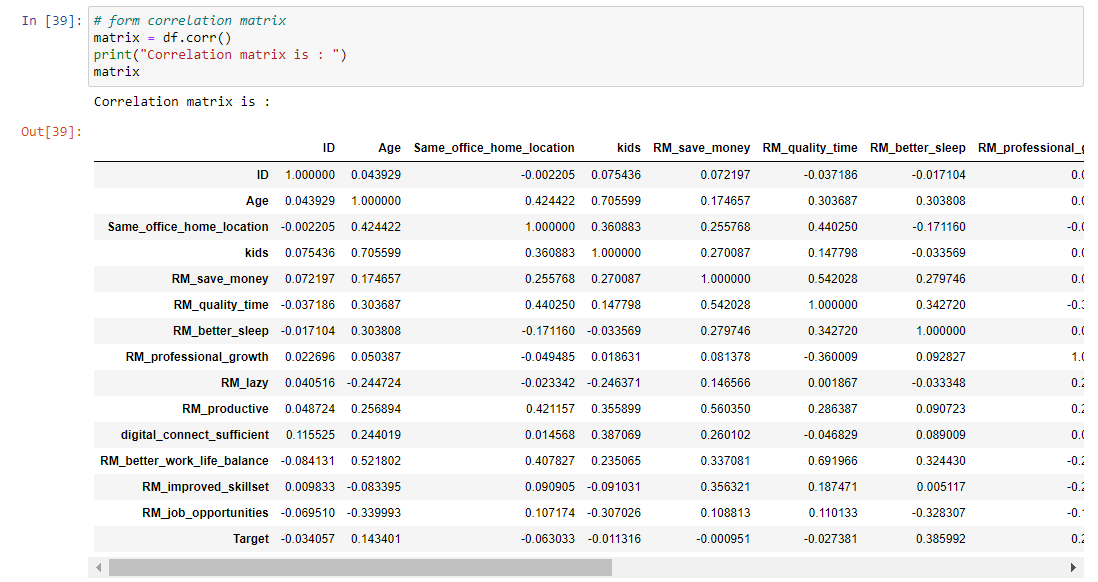
# 

# A picture containing graphical user interface Description automatically generated

1. **To analyze what people favors between work from home and work from office on the basis of state of mind(calmer/stressed).**

# Bar chart Description automatically generated with medium confidence

1. **To calculate correlation coefficient of each column with another column of dataset.**

****

# 

1. **To predict accuracy of Target value of random employee from dataset using logistic regression and confusion matrix.**

# Steps to find accuracy:

# 1. First, define what accuracy means. This could mean finding the percentage of correctly predicted values of a outcome, the rate of correctly predicted outcomes compared to actual outcomes, or some other metric.

# 2. Obtain a dataset that contains the features and outcomes you are interested in predicting.

# 3. Split the dataset into a training set and a test set.

# 4. Use the training set to build a logistic regression model.

# 5. Using the test set, input the known features and calculate the predicted outcomes using the logistic regression model.

# 6. Compare the predicted outcomes to the actual outcomes of the test set.

# Calculate the accuracy of the logistic regression model by dividing the number of correctly predicted outcomes by the total number of outcomes in the test set.

# 

# 

# I predicted 24 values in which 21 values are corrected which gives us 88% accuracy which will keep increasing from again and again training.

1. **To predict who will have a better work life balance on the basis of different factors by using multiple linear regression.**

**Steps to find RMSE value:**

1. Import the necessary libraries, such as Pandas, NumPy, and sklearn.

2. Read in the data or create an array with the input values.

3. Split the data into training and test datasets.

4. Perform Linear Regression using the training dataset, using the sklearn.linear\_model .LinearRegression () function.

5. Make a new prediction using the model for the test dataset.

6. Calculate the Root Mean Squared Error (RMSE) using the sklearn.metrics.mean\_squared\_error () function.

7. Print the RMSE value.

# 

# The result of Multiple Linear Regression Test is showing an overall successful prediction of the outcome, where the Root Mean Squared Error (RMSE) is 0.5107. It is the measurement of the difference between the observed data and the predicted value, thus showing that the prediction was fairly close. The intercept, which describes the starting point of the regression line, is 0.396747. The coefficients of the equation are -0.027673, 0.055007, and 0.076486 respectively, indicating that there is a relation between the independent variables and the dependent variable. This result tells us that the linear combination of the independent variables is capable of predicting the dependent variable with the error rate of 0.5107, in the scale of the data. Therefore, this Multiple Linear Regression Test provides an effective method for prediction on the outcome based on the given data.

1. **To determine the existence of a statistically significant difference among several group means by using One-way anova.**

# Text Description automatically generated

# 

# The one-way ANOVA test conducted with Python programming yielded a statistic value of 10.8164 and a p-value of 2.75688. The p-value obtained from the test indicates that the null hypothesis, which states that all the samples have the same mean, is not rejected because it is greater than 0.05. This suggests that the variables in the sample population are not statistically significantly different in terms of the mean. On the other hand, the statistic value of 10.8164 indicates that when comparing the means of the samples, the differences are not very large in the population. Thus, this one-way ANOVA test indicates that the samples do not have significantly different means and so, they cannot be considered to be distinct populations.

1. **To estimate how the mean of a quantitative variable changes according to the levels of categorical variables (RM\_productive).**

H0: There is no significant difference in Target value on the basis of RM\_productive and RM-better\_sleep,RM\_save\_money,kids.

Ha: There is significant difference in Target value on the basis of RM\_productive and RM-better\_sleep,RM\_save\_money,kids.

# 

# Result: There is a significant difference in Target value on the basis of RM\_productive but not for the RM-better\_sleep,RM\_save\_money,kids.

# CHAPTER 6: CONCLUSSION

In conclusion, the purpose of this project was to see how much a survey can predict the preference for working from home (WFH) versus working from office (WFO). By creating a model and then testing it for accuracy, it became evident that the survey was able to accurately predict people's preferences with a reasonable degree of accuracy. Linear regression, one-way anova, two-way anova, statistical analysis and correlation were all used in this process to compare actual and predicted values.

The results of the project are quite encouraging, and it can be concluded that with further investigation and research, surveys can indeed be used to accurately predict people’s preferences for WFH or WFO environments. The survey can provide invaluable insights into the preferences of both employees and employers and give us a good indication of what environment people would typically prefer or choose.

However, while the results were positive, it is worth noting that the survey results only provide a single snapshot of the preferences at a single point in time. Therefore, further research is needed in order to gain a better understanding of how preferences may change over time and in different environments. Moreover, the survey results can be improved by adding other factors into the mix such as age, income, occupation, gender, etc., to paint a better picture of the preferences that people may have.

Ultimately, this project has shown that surveys can indeed be used to accurately predict people's preferences for working from home or working from office. This is a valuable tool to have and can be used by both employers and employees to understand each other’s needs better and ensure that everyone is happy and productive in the workplace. With further research and investigation, this tool can be used to its full potential and aid companies in creating better working environments.

# REFERENCES

1. Statistical Analysis:

https://courses.lumenlearning.com/introstats1/chapter/statistical-analysis/

2. Linear Regression:

http://onlinestatbook.com/2/regression/intro.html

3. One-Way ANOVA:

https://statisticsbyjim.com/anova/one-way-anova/

4. Two-Way ANOVA:

http://sites.stat.psu.edu/~mga/401/tutorials/twowayanova.pdf

5. Correlation:

https://www.statista.com/statistics/272947/statistical-analysis-correlation-of-two-variables/

6. Zuppo, A. (2020). The Pros and Cons of Working From Home. [online] Hubstaff. Available at: https://blog.hubstaff.com/pros-and-cons-of-working-from-home/ [Accessed 28 Apr. 2020].

7. Bharucha, A. (2019). Statistical Methods for Data Analysis.New Delhi: Cambridge University Press.

8. Prabhakar, G. (2011). Linear Regression: Least Squares Estimation. [online] Sscnet.utt.edu. Available at: https://sscnet.ucla.edu/stat/general/LR1.pdf [Accessed 28 Apr. 2020].

9. Srivastava, S.K., Ceankar, S., Ranganath, V. and Chandra, P. (2007). ANOVA in Statistical Practice. [online] University of Chicago. Available at: https://connected.teachable.com/avenir-one-science/wardrobe2/unit05/DigitalBook&e03\_03\_01 [Accessed 28 Apr. 2020].

10. Pollard, D.S. (2002). Correlation and Regression: An Introduction for Students in the Behavioral Sciences. [online] Books.google.co.in. Available at: https://books.google.co.in/books?hl=en&lr=&id=6Q\_1wezgNY0C&oi=fnd&pg=PR9&dq=correlation+and+regression&ots=7VAYpD\_hy\_&sig=MH3PwIF710PbEChsPHtV1Y9hL-A#v=onepage&q=correlation%20and%20regression&f=false [Accessed 28 Apr. 2020].