



University of Findlay

# PREDICTING THE PREFERENCE BETWEEN ONSITE vs REMOTE WORKING

Nahom Gebremedhin

Nov. 26, 2023

# Abstract

- In the wake of the COVID-19 pandemic, the global workforce witnessed a major shift towards remote work arrangements. Since then, the number of people working remotely has increased dramatically.
- This capstone project aims to explore and compare employees' preferences between working from home and remotely.
- It examines the key factors influencing employees' choices between the two work domains, analyzing the various factors that impact their decisions. Factors such as productivity, work-life balance, job satisfaction, and the overall impact on their personal and professional lives.

# Project Description

- The research project aims to predict individuals' preferences between working from home and working onsite. With the rise of remote work, it has become increasingly important to understand the factors that influence individuals' preferences for working arrangements.
- By examining the predictors of these preferences, such as demographic and job-related factors, the study aims to provide insights for employers and policymakers on how to optimize remote work arrangements to meet the needs and preferences of their workforce.
- To create a work environment that suits an employee both personally and professionally for job satisfaction and improved productivity based on the insights derived from the analysis.

# Beneficiaries

- This is basically intended to assist companies, in optimizing productivity, which run their businesses in different work domains such as:
  - Onsite
  - Remote
  - Hybrid

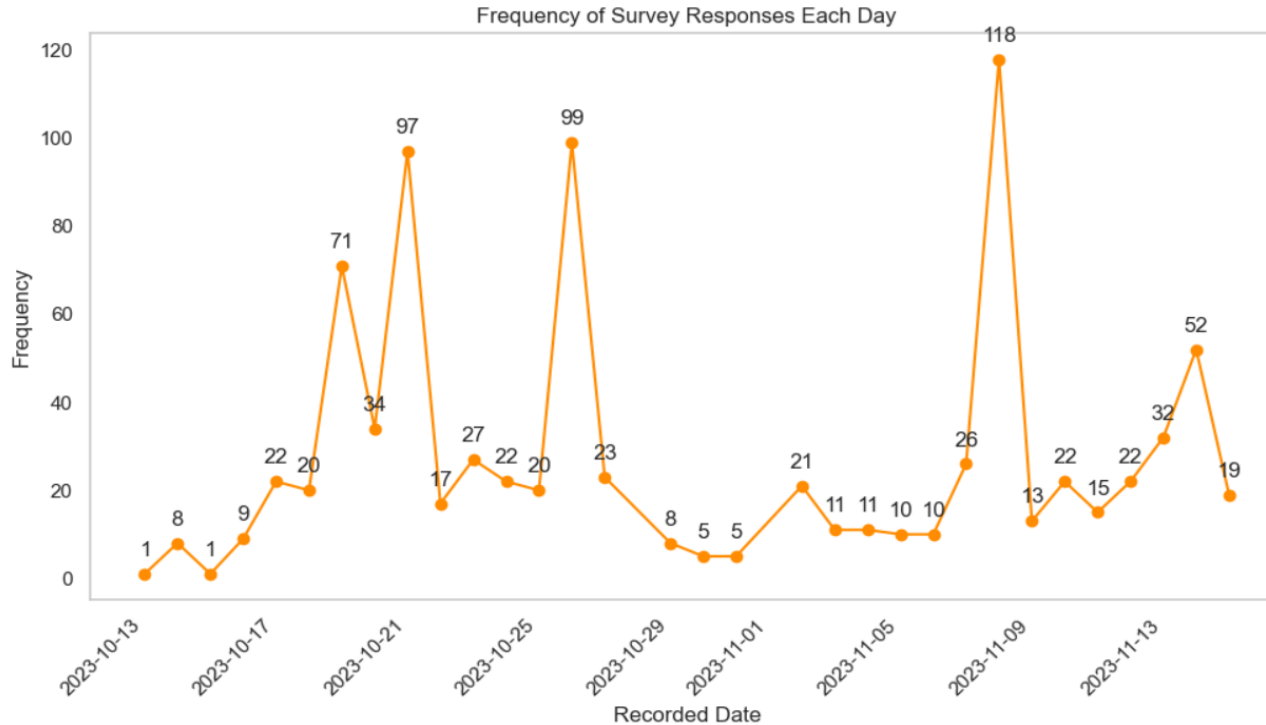
# Research Questions

- What is an employee's preference in the work domain, onsite, remote, or hybrid working?
- What factors influence employees' preferences between working from home and working onsite?
- Are there demographic factors (age, experience, etc.) that correlate with specific work preferences?

# Data Set

- The study used online survey platforms, mostly Qualtrics, to gather data from a wide range of employees.
- The data collection process used two primary sources: Amazon Mechanical Turk (MTurk), a crowdsourcing platform to gather responses from a diverse set of individuals, and a LinkedIn-shared survey link, to collect responses from individuals with different professional backgrounds.
- In total, 872 responses were collected and used for analysis.
- The survey had been made available for a month.

# Data Set (continued)



- This graph shows the number of responses collected from the online survey tool.
- The peaks are responses from MTurk.

# Data Set and Research Questions Linkage

- The dataset includes variables such as age, gender, occupation, remote work productivity, and remote work fit related to the research questions. These variables provide a view of employees' characteristics and their preferences.
- By examining correlations between various factors and employees' preferences, the research questions were partially addressed.



# Data Cleaning and Preprocessing

- After the data were collected, they were cleaned and preprocessed for further analysis.
  - The data cleaning and preprocessing included recoding categorical variables.
  - There was no need to impute missing values.
- In addition, charts were produced to give an overview of the data.

# Revised Data Set

	A	B	C	D	E	F
1	Random ID	1449	1510	1465	1784	1733
2	What is your age?	20	54	25	32	24
3	What is your gender?	Female	Female	Female	Male	Female
4	Which of the following best describes your current occupation?	Education	Business	Business	Business	Business
5	Do you live in the same city as your office?	No	No	Yes	Yes	Yes
6	If yes, how far is your workplace from your residence?			1-5 miles	6-10 miles	1-5 miles
7	How many years of experience do you have in your current job role or profession?	2-3 years	2-3 years	2-3 years	7-10 years	2-3 years
8	How many children, under 12, do you have living with you?	None	One	One	Two	Two
9	On average, how many hours do you work per day?	Less than 8 hours	Less than 8 hours	8 hours	More than 8 hours	Less than 8 hours
10	On average, how many hours do you spend with family/friends per day?	1 hour	1 hour	1 hour	3 or more hours	1 hour
11	On average, how many hours of sleep do you get per day?	Less than 6 hours	6-8 hours	6-8 hours	6-8 hours	Less than 6 hours
12	Do you engage in remote work, even if only occasionally?	Yes	No	Yes	Yes	Yes
13	How stressful is your job since your remote work began?	Little or no stress		Extremely stressful	Moderately stressful	Little or no stress
14	Do you feel working from home has affected your professional growth?	Yes		Yes	Yes	Yes
15	How productive are you with working remotely?	Less productive		Extremely productive	Very productive	Productive
16	Do you feel you are able to save money with remote work?	No		Yes	Yes	Yes
17	How satisfied are you with your current work-life balance?	Neither satisfied nor	Neither satisfied nor	Extremely satisfied	Satisfied	Satisfied
18	How important do you think meeting with co-workers in person is compared to online?	Slightly important	Important	Extremely important	Extremely important	Important
19	Do you think there are more job opportunities with remote work?	Yes	Yes	Yes	Yes	Yes
20	Would you take a pay cut to work remotely?	Yes	No	Yes	Yes	Yes
21	If yes, how much of a pay cut would you take?	Less than 5%		5-10%	More than 15%	5-10%
22	In your opinion, how well does remote work align with your work preferences?	Not very well	Not very well	Very well	Extremely well	Neutral
23	Ideally, with a five-day work week, on average how many days a week would you prefer to work remote (with other days being on site)?	2 days	2 days	1 day	5 days	2 days
24	Is there anything else that you wish to share related to remote work?	I can save my time of travelling	better balance home and work life	GOOD	No comments	

# Analytic Techniques/Models

- Some common analytic techniques and models used:
  - Descriptive Statistics: this includes measures such as mean, median, min, max, standard deviation, and count.
  - Correlation Analysis: to assess the strength and direction of the linear relationship between two continuous variables.
  - Chi-Square Test: to determine if there is a significant association between two categorical variables.
- Analytics tools and statistical software used were:
  - MS Excel.
  - Python.

# Feature Selection

- The next step was “feature selection”. Based on the chi-squared test results, features highly related to the target variable were selected for training predictive models.
- The following are some of the independent variables with a significant relationship to the target variable, days to work remotely.

# Chi-Squared Test Results

Chi-squared test results:

	Feature	Chi-squared statistic	P-value
0	Age	274.561201	0.0072659218
1	Children_under12	57.190204	0.0000007619
2	Experience	94.152099	0.0000000000
3	Gender	18.555766	0.0023247796
4	Hours_of_sleep	6.256370	0.7932838159
5	Hours_worked	49.080533	0.0000003937
6	Inperson_meeting_importance	118.666891	0.0000000000
7	Job_stress(Remote)	65.534408	0.0000000275
8	Leisure_time_per_day	61.162875	0.0000001590
9	Location	20.804078	0.0008820807
10	Occupation	67.269067	0.0000005052
11	Pay_cut_Amount	109.056812	0.0000000000
12	Pay_cut_for_remotework	16.401559	0.0057863916
13	Profession_impact	48.169942	0.0000005779
14	Random ID	2874.640779	0.2970681183
15	Remote_Work_Opportunities	21.008999	0.0008068947
16	Remote_work_fit	240.126123	0.0000000000
17	Remote_work_frequency	21.261049	0.0007230305
18	Remotework_Productivity	160.250663	0.0000000000
19	Remotework_Savings	24.537837	0.0062938744
20	Work-Life balance satisfaction	81.731458	0.0000000020
21	Workplace distance	59.714545	0.0000078836

# Example of a Feature Selection

Cross-tabulation table:

Days_to_work_remotely	0	1	2	3	4	5
Inperson_meeting_importance						
1	1	4	13	6	5	27
2	4	24	34	47	23	17
3	2	38	79	119	83	32
4	4	11	62	78	69	30
5	1	7	6	11	12	22

Chi-square value: 118.66689131951713

P-value: 5.031701347973983e-16

Hypotheses:

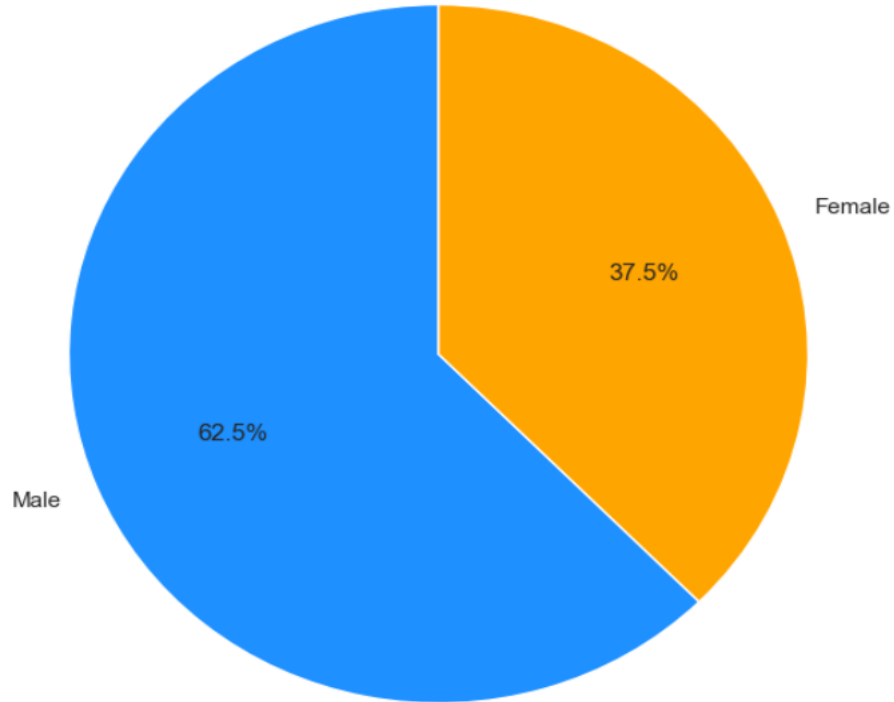
$H_0$ : There is no association between In-person Meeting Importance and Days to Work Remotely.

$H_a$ : There is an association between In-person Meeting Importance and Days to Work Remotely.

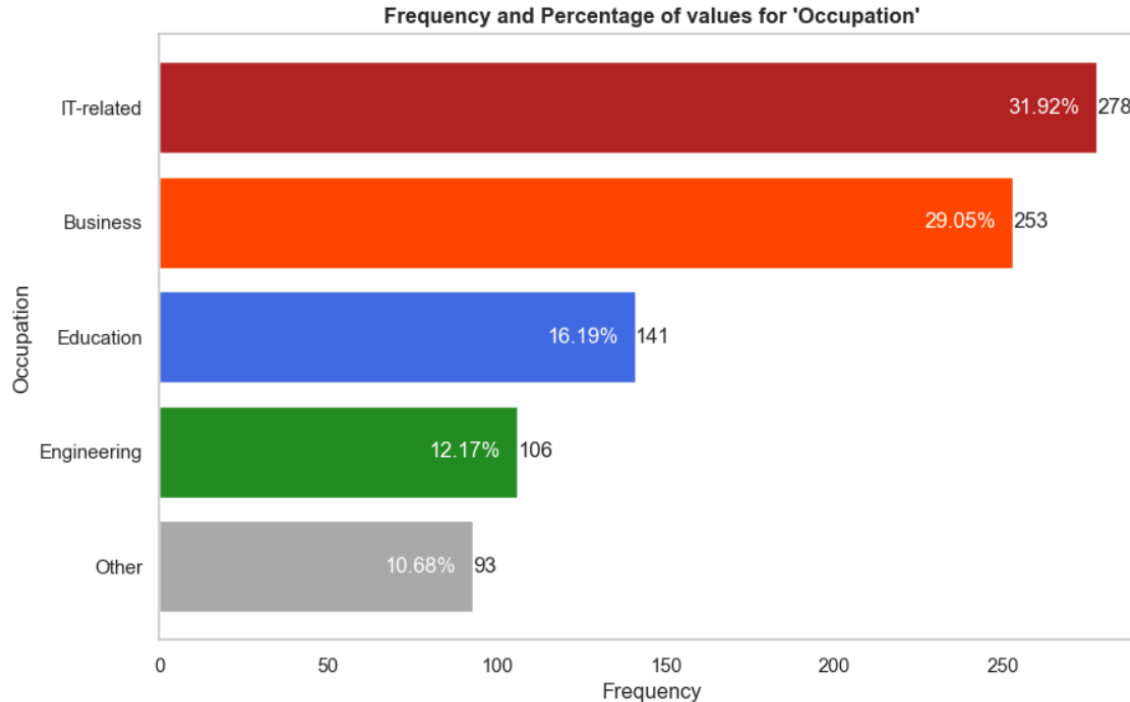
Conclusion: Reject the null hypothesis. There is evidence of an association.

# Descriptive Statistics and Visualizations

Gender Distribution



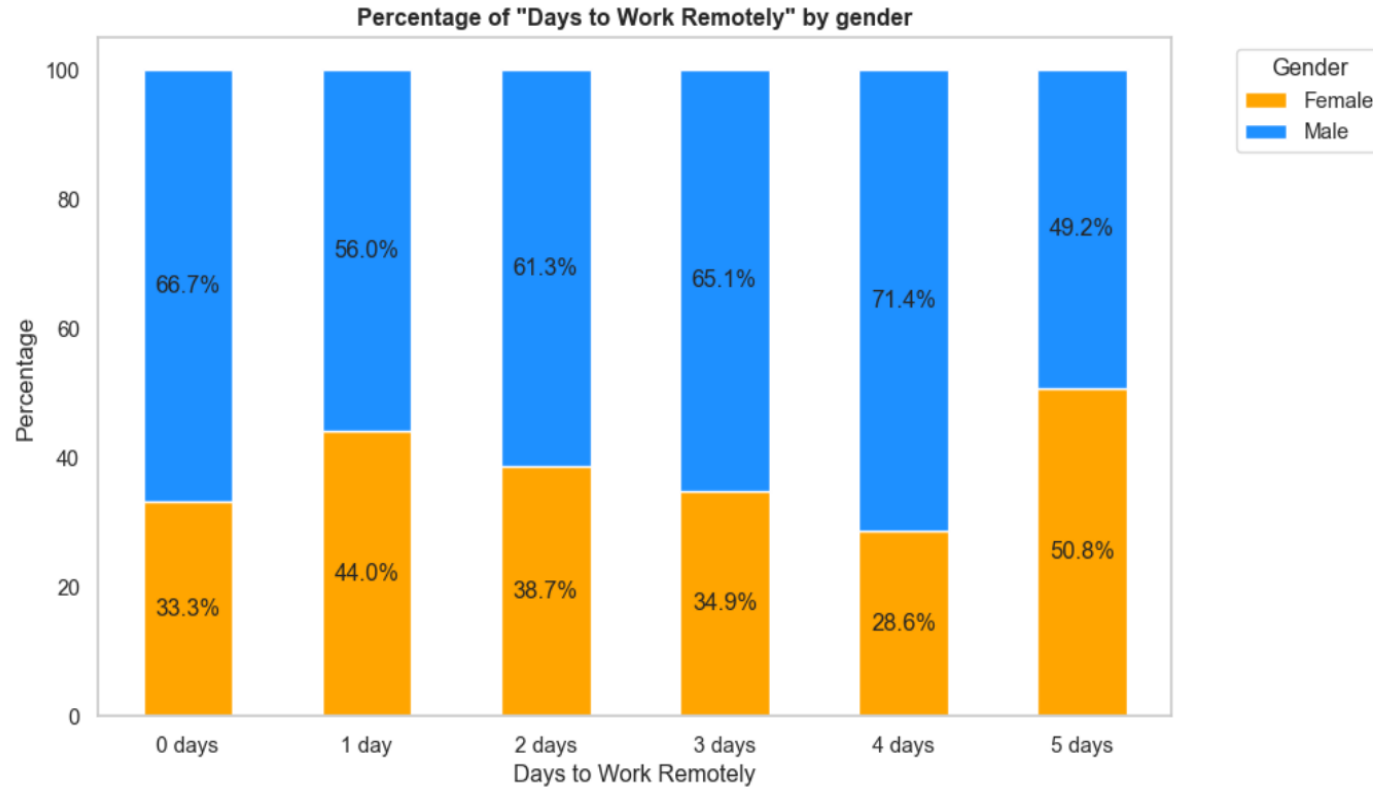
# Descriptive Statistics and Visualizations – cont.



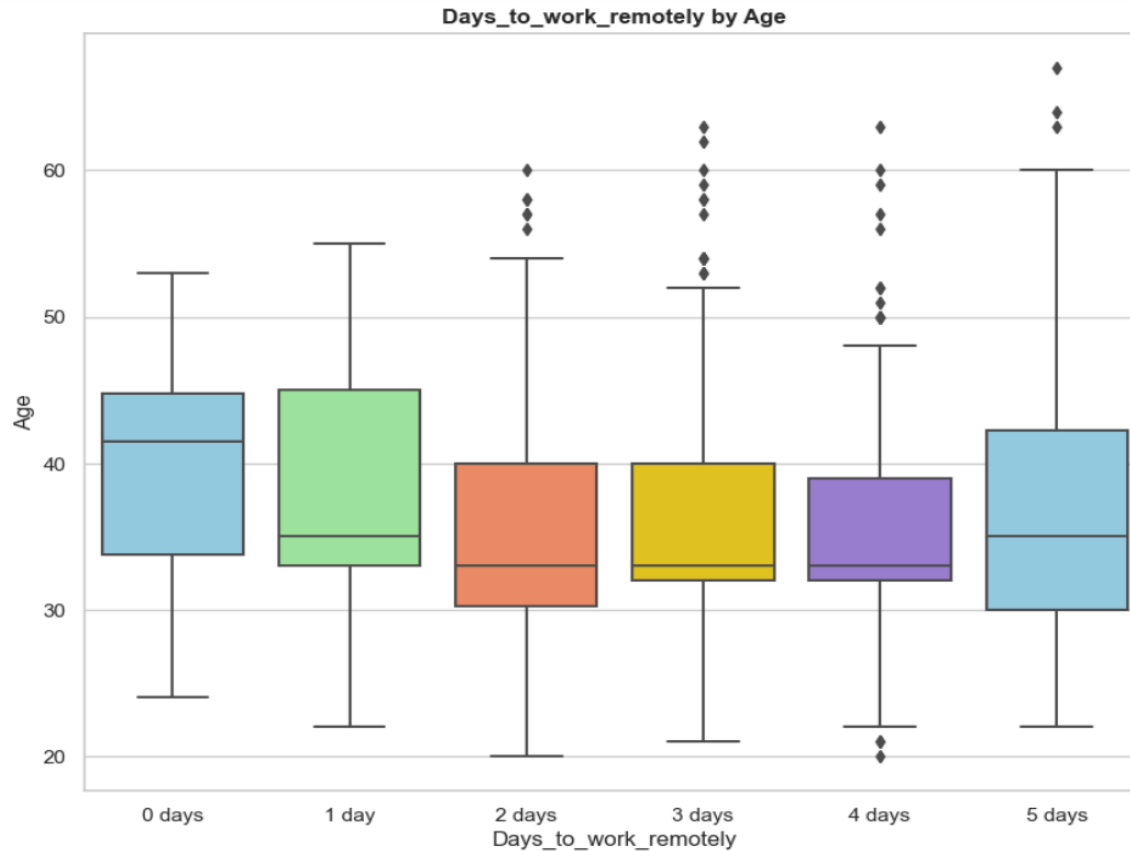
- The above chart shows distribution of survey participants by profession with the highest number of participants being in IT-related occupation.



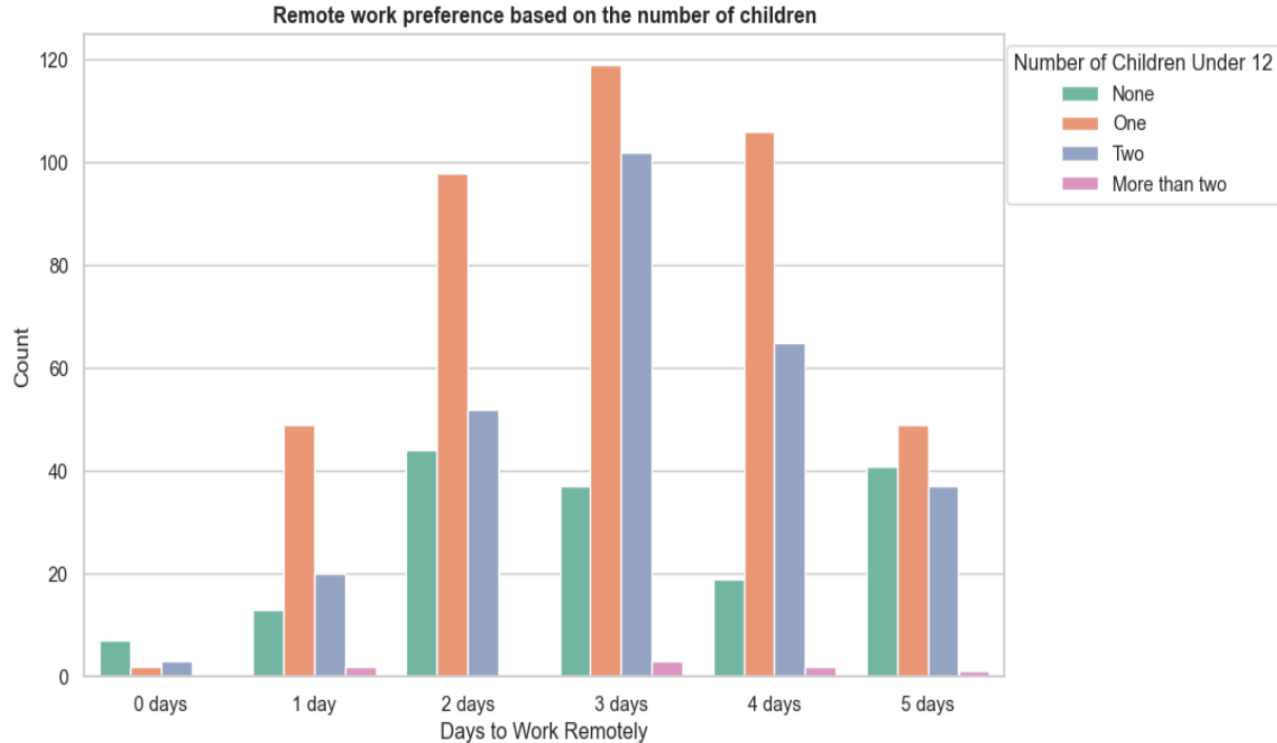
# Descriptive Statistics and Visualizations – cont.



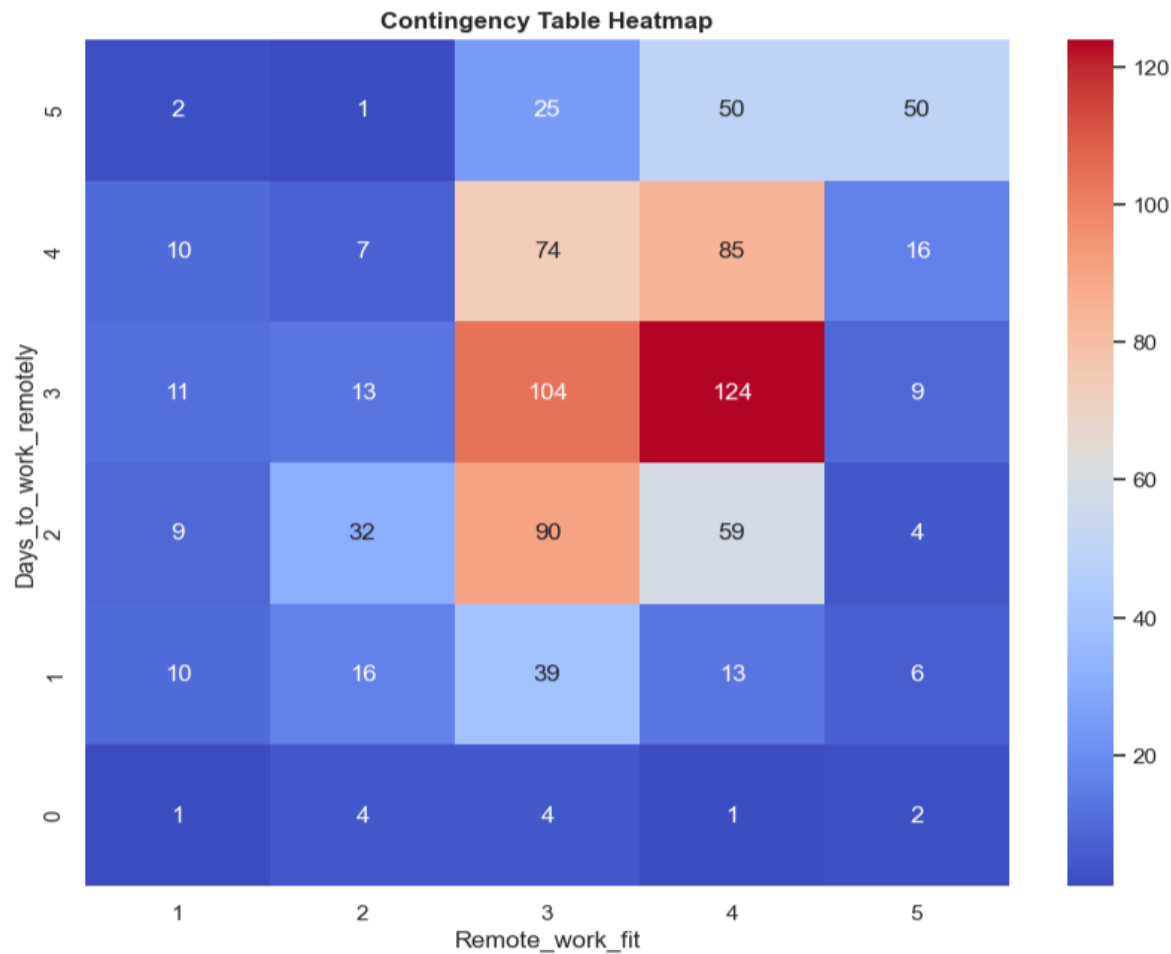
# Descriptive Statistics and Visualizations – cont.



# Descriptive Statistics and Visualizations – cont.



- The number of children an employee has doesn't seem to have impact on their remote work-days preference.



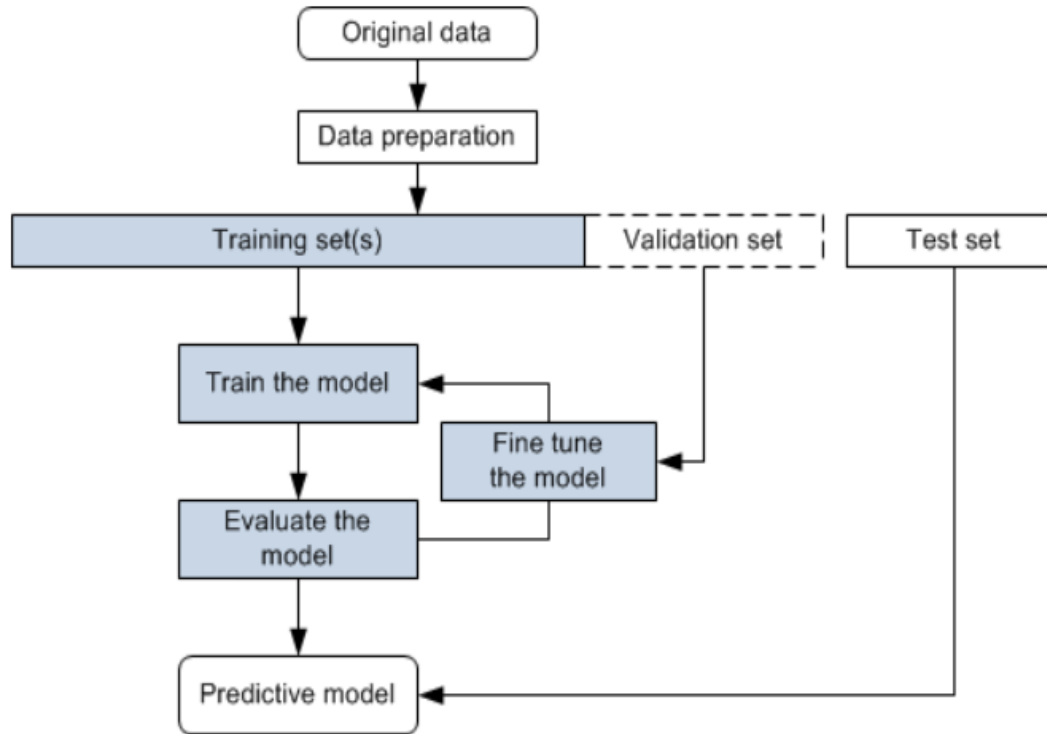
# Contingency Table Interpretation

- As we move from “Remote Work Fit” 1 to “Remote Work Fit” 5, there is a general trend of increase in the number of days people work remotely.
- Across all “Remote\_work\_fit” levels, there are very few instances of people working 0 days remotely. This suggests that a majority of people in the dataset tend to work at least 1 day remotely. Even those who think remote work is not a good fit for them.
- “Days\_to\_work\_remotely” 3 stands out as having relatively high frequencies across all categories of “Remote\_work\_fit”.

# Methodology

- After feature selection, the dataset was divided into
  - Training, 50% of total
  - Validation, 20% and
  - Testing sets, 30%
- Several classification models were trained for predictive analysis.
- Some of the predictive modeling techniques implemented are:
  - Ordinal logistic regression
  - Random forest
  - Decision tree classifier
  - Ada boost
  - Xgboost classifier as well as
  - K-means clustering analysis

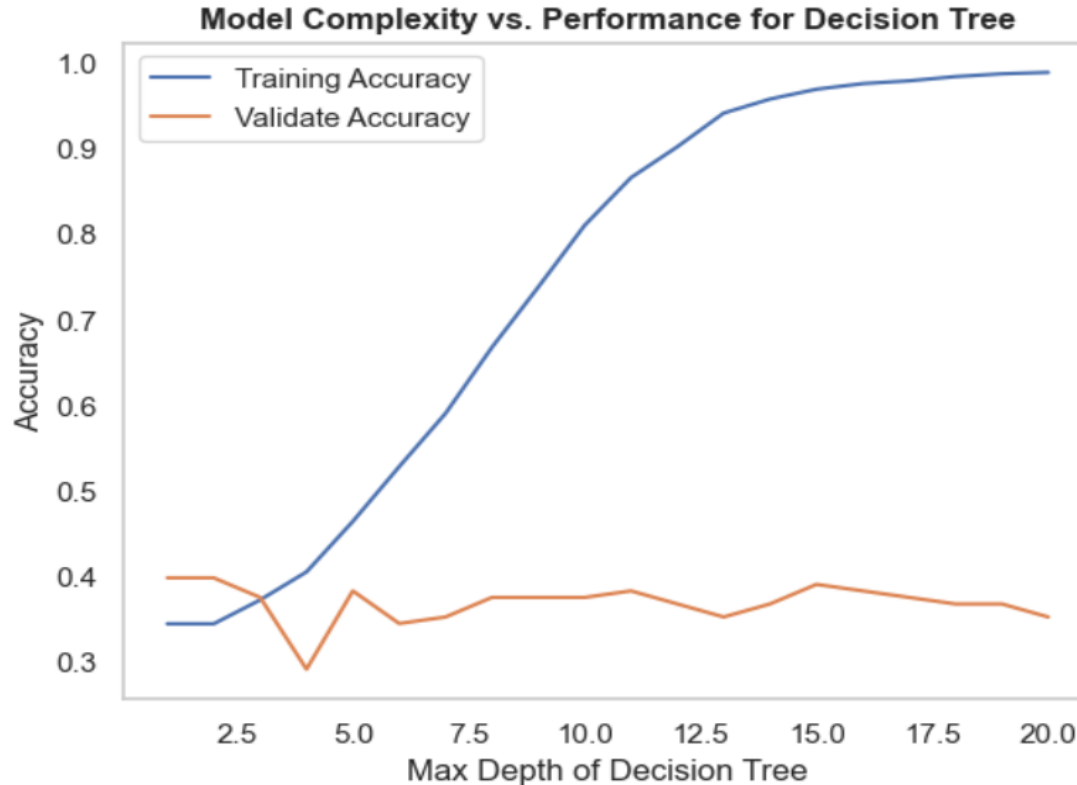
# Model Training Flowchart



Reference:

Workflow and data preparation. (n.d.-b). <https://www.theobjects.com/dragonfly/dfhelp/2020-1/Content/Artificial%20Intelligence/Deep%20Learning%20Tool/Workflow%20and%20Data%20Preparation.htm>

# Advanced Analysis



- A maximum depth of three was used for better performance of the model.



# Decision Tree Results

Accuracy: 0.3816793893129771

Precision: 0.48392508432451625

Recall: 0.3816793893129771

F1-Score: 0.39244557767364563

Confusion matrix:

```
[[ 0  0  0  1  0  0]
 [ 0  2  5  2  2  0]
 [ 0  1 13 11  8  0]
 [ 0  0 10 18 12  0]
 [ 1  0  5  9  9  0]
 [ 0  0  3  2  9  8]]
```

# Random Forest Results

Accuracy: 0.3969465648854962

Precision: 0.4373645014866389

Recall: 0.3969465648854962

F1-Score: 0.39950736351626615

Confusion matrix:

```
[[ 0  0  0  1  0  0]
 [ 0  3  5  2  1  0]
 [ 0  3 15 10  5  0]
 [ 0  0 11 19  9  1]
 [ 0  1  0 15  7  1]
 [ 0  0  4  5  5  8]]
```

# XGBoost Results

Accuracy: 0.35

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.33	0.27	0.30	11
2	0.34	0.33	0.34	33
3	0.34	0.38	0.36	40
4	0.22	0.29	0.25	24
5	0.71	0.45	0.56	22
accuracy			0.35	131
macro avg	0.33	0.29	0.30	131
weighted avg	0.38	0.35	0.36	131

Confusion Matrix:

```
[[ 0  0  0  1  0  0]
 [ 0  3  4  1  3  0]
 [ 0  4 11 10  6  2]
 [ 0  2 11 15 11  1]
 [ 0  0  4 12  7  1]
 [ 0  0  2  5  5 10]]
```

# AdaBoost Results

Accuracy: 0.35

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.29	0.18	0.22	11
2	0.43	0.27	0.33	33
3	0.32	0.28	0.30	40
4	0.27	0.54	0.36	24
5	0.55	0.50	0.52	22
accuracy			0.35	131
macro avg	0.31	0.30	0.29	131
weighted avg	0.37	0.35	0.35	131

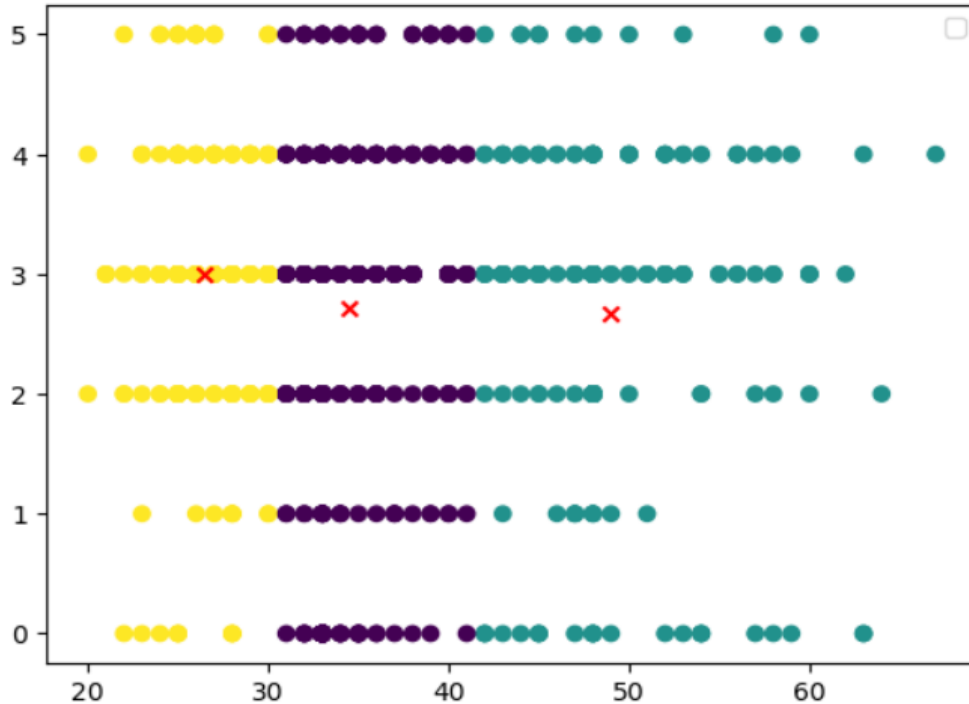
Confusion Matrix:

```
[[ 0  0  0  1  0  0]
 [ 0  3  4  1  3  0]
 [ 0  4 11 10  6  2]
 [ 0  2 11 15 11  1]
 [ 0  0  4 12  7  1]
 [ 0  0  2  5  5 10]]
```

# Performance

- The performance of each model was assessed using the validation set.
- Various evaluation metrics, such as precision, f-1 score, with a focus on accuracy, were employed to understand the strengths and weaknesses of each model.
- Most of the models have low level of performance.

# K-Means Cluster Analysis



- The variables were clustered into three based on the preference for the 'Days to work remotely' which ranges from 0 days to 5 days per week.

# K-Means Cluster Analysis Results

	0	1	2
<b>Age</b>	34.474684	48.913043	26.421053
<b>Remotework_Productivity</b>	2.717300	2.666667	2.994737
<b>Pay_cut_Amount</b>	1.582278	1.454106	1.742105
<b>Workplace distance</b>	2.052743	1.951691	2.100000
<b>Remote_work_fit</b>	3.415612	3.376812	3.378947
<b>Days_to_work_remotely</b>	3.046414	2.985507	3.163158

- Different k values were used but the k=3 is used as the data can be adequately represented by three distinct groups.

# Roadblocks & Challenges

- Unable to collect as much data as desired.
- Obtaining a representative sample of the population of interest.
- Identifying and recruiting individuals from diverse professional backgrounds.
- Obtaining a predictive model with high performance on the data.



# Limitations

- The data collection process was such that generalizing to the desired population is limited.
- The survey captures preferences at a specific point in time, but those preferences may not be stable. So, it's not a one-time work.
- The quality of participants' home working environments were not considered.
  - It can significantly impact their preferences or desire for remote work. Factors such as available space, family distractions, or the lack of a dedicated home office space.
- The survey didn't address issues related to technology accessibility.
  - Employees without access to reliable internet connections or necessary technology have not been represented.

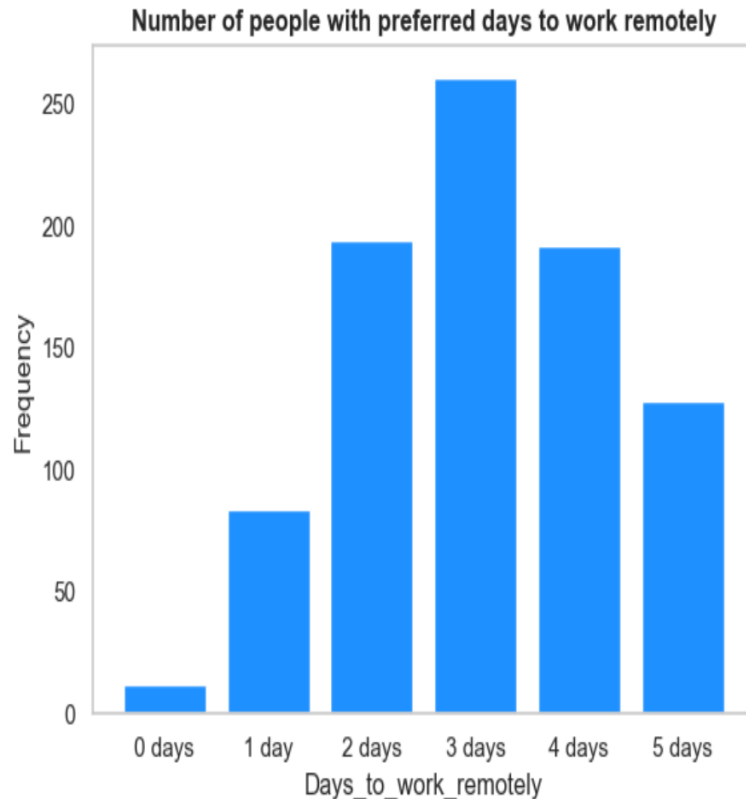
# Next Steps

- Incorporating questions related to context of participants' work environments or the specific conditions that influence their preferences.
- Including open-ended questions or interviews to understand the reasoning behind their preferences.
  - Text analysis techniques can be applied to extract insights and patterns from the text data.
- Adding personality trait questions to see if they influence the choice between working from home and onsite?
  - For example, if the person is introvert or extrovert.
- Conduct additional K-means clustering analysis.

# Result

- The research indicates a widespread preference for remote work with higher preference for hybrid work, with highest number of people preferring 3 days a week. So, future analysis of remote vs onsite work environment should include a hybrid option.
- As some mentioned in the comment section, this preference is mainly because it offers flexibility, saves time by avoiding daily commutes, and helps cut costs related to office space and commuting associated with traditional office-related expenses.

# Result (continued)



## Is there anything else that you wish to share related to remote work?

better balance home and work life

I can save my time of travelling.

Rather than spending an hour a day in commute, I could use this time to plan and execute my work.  
its peaceful and spend our time with our family.

I want to work at home so I can be more productive and use all my available time to focus on my job.

I feel freedom and flexibility in remote work..

I wish i could work remotely forever.

Best thing that has happened to me in >30 years of working.

I love working from home with flexible hours.

Remote work does so much to cut down on my commuting costs.

I like working from home but it would be nice if I occasionally could go into an office somewhere and see my coworkers.

Another thing influencing my preference for remote work is the fact that the quality and safety of public transport also seems to be declining (for example, the New York City MTA system).

It is a type of flexible working arrangement that allows an employee to work from remote location outside of offices

It's better in terms of saving money from driving to work, buying food during lunch and it allows you to take breaks the way you want and how long you want as long as you get your job done. You have more freedom and you feel like you're responsible for your own time.

I like having a flexible job, it gives me relief knowing that I can focus on non-work related issues while knowing that I won't have to sacrifice my job for them.

Remote work is good only if you have quite and stress free environment.

# Conclusion

- In conclusion, companies should be ready to adapt to this work environment by aligning their strategies with the changing expectations and demands of the workforce.
- This adaptation is crucial for businesses aiming to remain competitive and responsive to the changing dynamics of the work environment.
- In the coming years, remote work opportunities will have to meet the demands and expectations of the future workforce.

# Thank You!!