**GROUP PROJECT PROPOSAL**

**ADTA 5130**

DR. YIXUN XING

**GROUP – 16**

**Group Members:**

Harini Chaturvedula (11522208)

Prakash Nanne (11515064)

Udayraj Suthapalli (11518446)

Venkat Akshay Reddy Jaggavarapu (11514134)

**INTRODUCTION:**

The **Health Insurance and Hours Worked by Wives Dataset** is being used to understand the Health insurance for the people who are working and how their job is related to their insurance policy. The objective of the project is to draw a new policy based on the conclusion drawn from the research. It will help to identify the employees who are eligible for health insurance from the company but still they didn’t get it and also to identify who don’t require it but still they have it. The additional data of wives working will also help us to draw better conclusion whether wives get HI from husbands work or their work and vice versa.

**OBJECTIVE:**

Trying to find out if HI is most likely to dependent on salary or working hours. The minimum full-time working hours for an employee in United states is 40 hours per week.

H0: Health insurance will be sponsored by company if >= 40 hours per week

HA: Health insurance will be sponsored by company if < 40 hours per week.

**DATASET:**

We have used the health insurance and hours worked by Wives dataset for this research and the dataset contains a total of 13 variables/columns which are

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S no** | **Variable name** | **Type** | **Variable type** | **Description of Variable** |
| 1 | Whrswk | Num | Continuous | No. of hours worked by wife per week |
| 2 | Hhi | Char | Dichotomous | Is the wife health insurance covered by husband’s HI |
| 3 | Whi | Char | Dichotomous | Is the wife health insurance covered with her job |
| 4 | Hhi2 | Char | Dichotomous | Is the husband health insurance covered with his job |
| 5 | Education | Char | Ordinal | No of years completed by education |
| 6 | Race | Char | Categorical | Type of race |
| 7 | Hispanic | Char | Dichotomous | Whether the employee is Hispanic or not |
| 8 | Experience | Num | Continuous | Number of years of experience the employee has |
| 9 | Kidslt6 | Num | Continuous | Number of kids whose age is less than 6 |
| 10 | Kids618 | Num | Continuous | Number of kids whose age is between 6 and 18 including 6 and 18. |
| 11 | Husby | Num | Continuous | Income of the husband in thousand dollars per year |
| 12 | Region | Char | Categorical | Belongs to which part of US regions. |
| 13 | Wght | Num | Continuous | No. of samples drawn to conclude hypothesis |

We have the data of 22,272 data which would be very much enough to conduct our experiment and draw conclusions.

**HYPOTHESIS STATEMENTS:**

* Is the husband’s income related to the health insurance of the wife or husband?
* Other factors like No of kids, experience, region, education how these are related to the health insurance of the family and income.
* Is it possible that higher working hours from the wives can get their health insurance from the company?
* Does the health insurance of the husband vary from the wives working hours if the husband annual income is less?

**TEAM WORK AND CONTRIBUTION:**

We are following all the steps of the CRISP-DM framework and trying to get the conclusions from the dataset. First, we will do the data wrangling and data cleaning to start our analysis.

1. Harini Chaturvedula: - Will work on the Hypothesis 1 which will get the correlation between the husband income and the health insurance of the both husband and wife.
2. Prakash Manne: - Will work on the Hypothesis 2 to get some useful insights from the different types of attributes.
3. Udayraj Suthapalli: - Will work on the Hypothesis 3 which will decide whether Wives health insurance will get covered from husband job or her own job.
4. Venkat Akshay Reddy Jaggavarapu: - Will work on Hypothesis 4 whether the wives working hours help the family to get the health insurance from the company she is working on.

We want to learn all the steps of CRISP-DM so instead of working on different step we thought of taking different problem statements and work on each of them using all the steps of the CRISP-DM.

At last we will collaborate regularly if we feel that one of our hypothesis statements is dependent on any of the other statement.

**COLLABORATION PLAN:**

|  |  |  |
| --- | --- | --- |
| Date | Tasks | Mode of Communication |
| 11/12/2021 | Project Proposal | Microsoft Teams |
| 11/14/2021 | Data Exploration | Microsoft Teams |
| 11/19/2021 | Data pre-processing, includes data wrangling and cleaning | In person |
| 11/24/2021 | Data Visualization | Microsoft Teams |
| 11/28/2021 | Statistical Analysis | In person |
| 12/01/2021 | Documentation | Microsoft Teams |
| 12/02/2021 | Conclusions | Microsoft Teams |
| 12/03/2021 | Final documentation | Microsoft Teams |
| 12/05/2021 | Review | Microsoft Teams |