Final Project

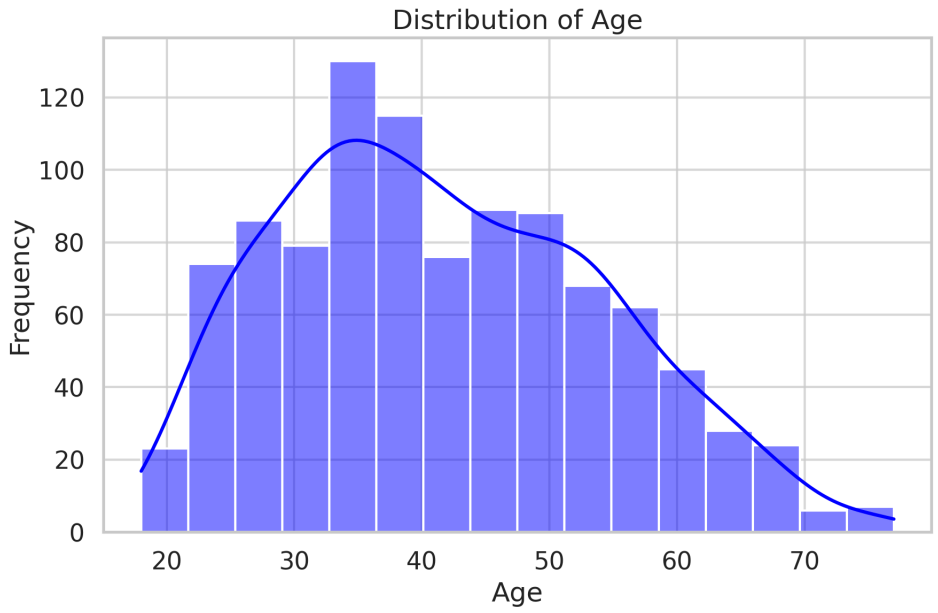
Caleb Morse
MIS445
07/29/2024

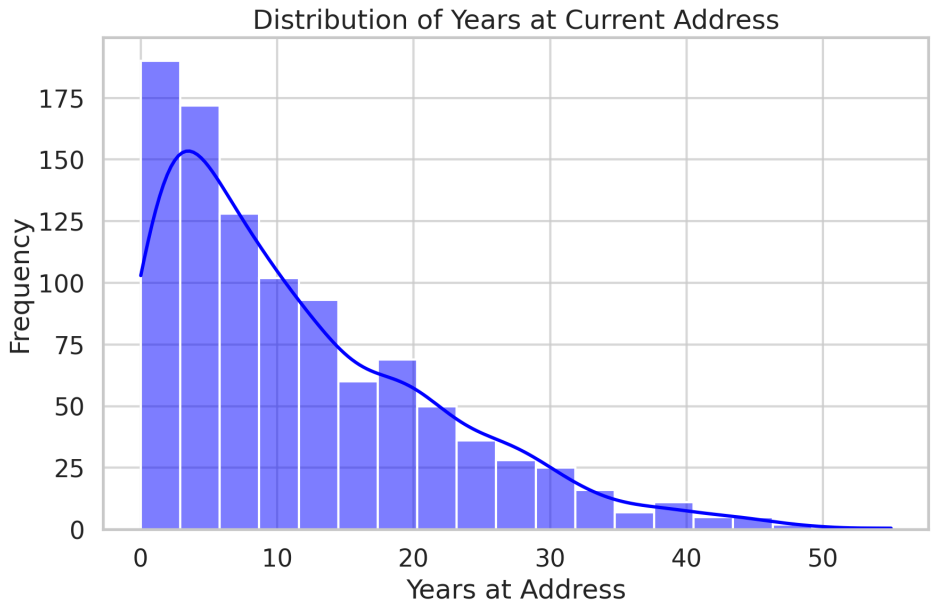
The purpose of this report is to analyze the Telco dataset to understand the characteristics of the company's customers and to predict their income and churn status. This analysis aims to help the telecommunications company reduce its dependency on external vendors for income data by developing an internal predictive model. The report includes sample characteristics, multiple linear regression analysis, hypothesis testing, and interpretation of the results.

Descriptive Statistics

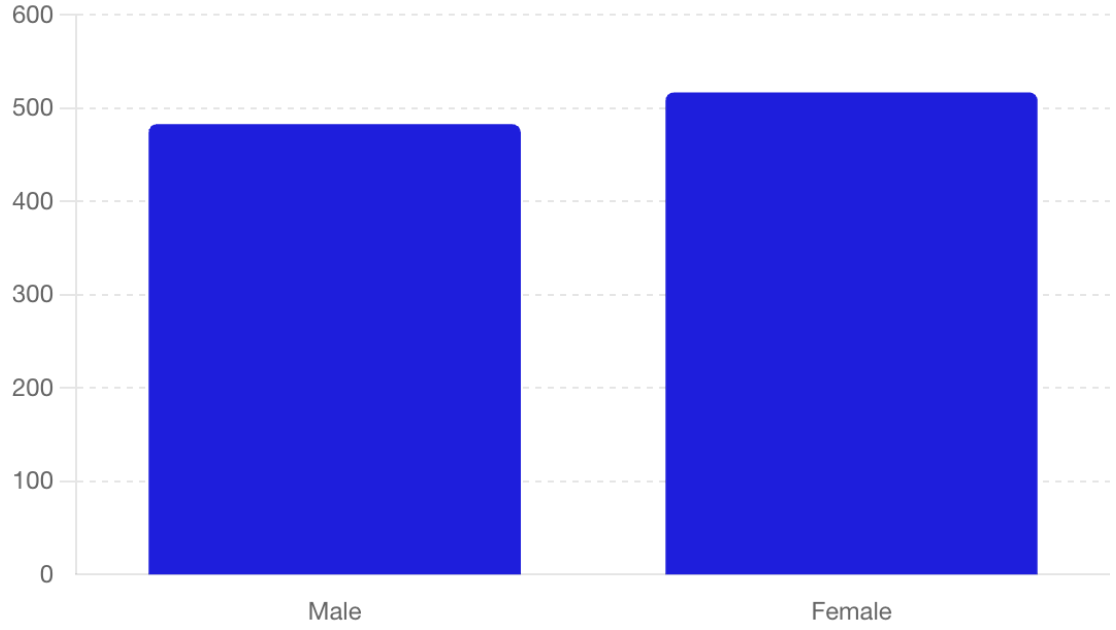
I will calculate and visualize the descriptive statistics for the selected variables.

1. **Age**
2. **Years at Current Address**
3. **Gender**
4. **Level of Education**
5. **Income**
6. **Marital Status**
7. **Region**
8. **Custcat**
9. **Churn**

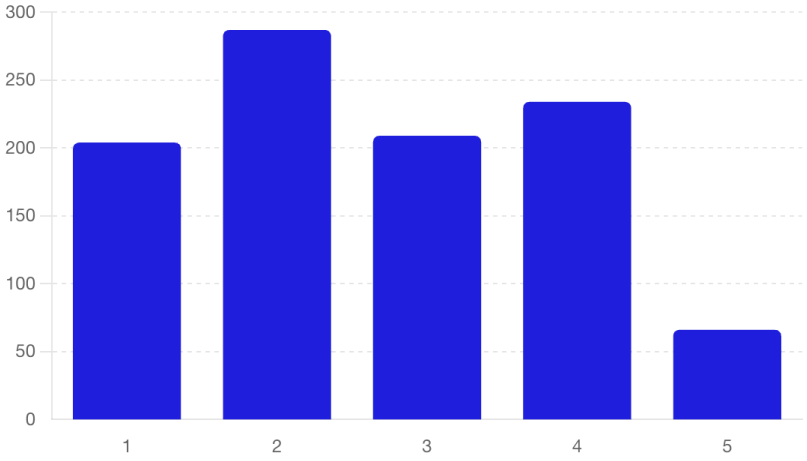


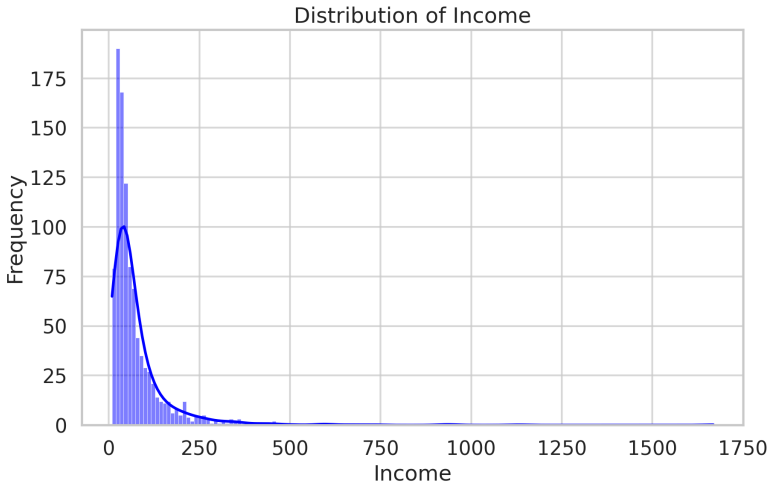


Distribution by Gender

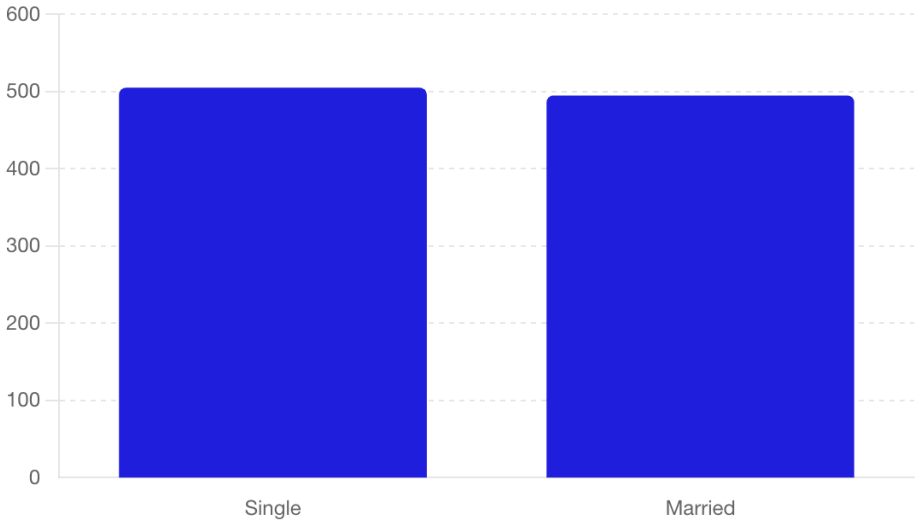


By level of education

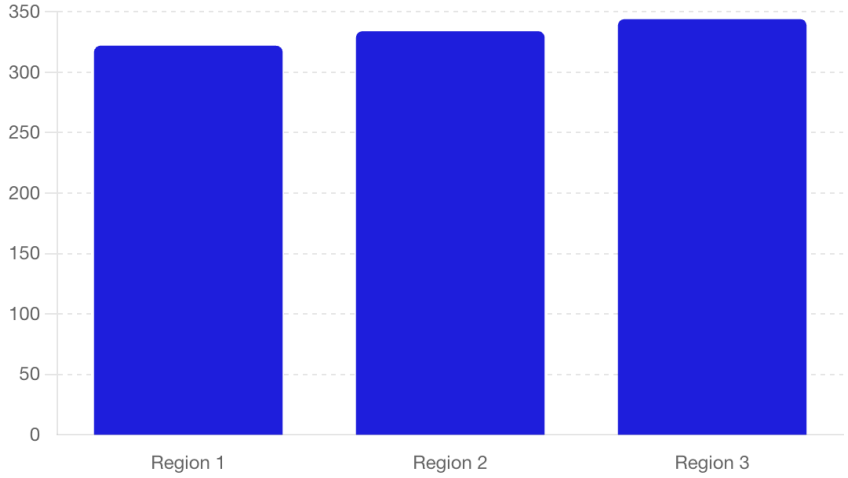




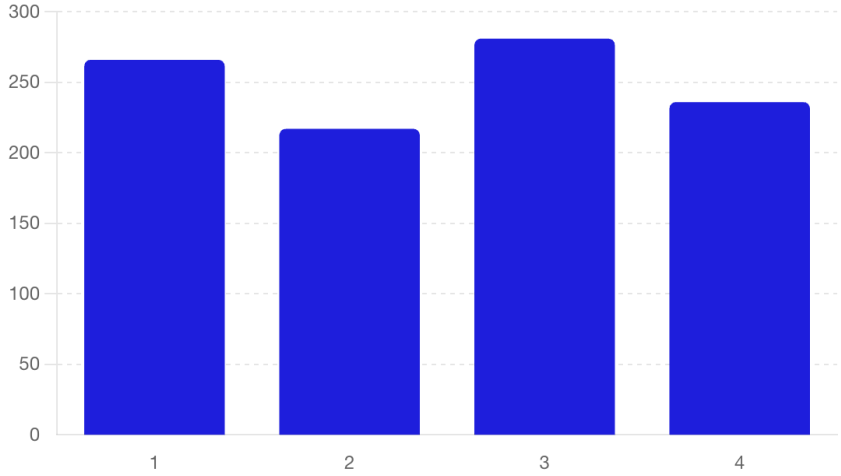
Marital status



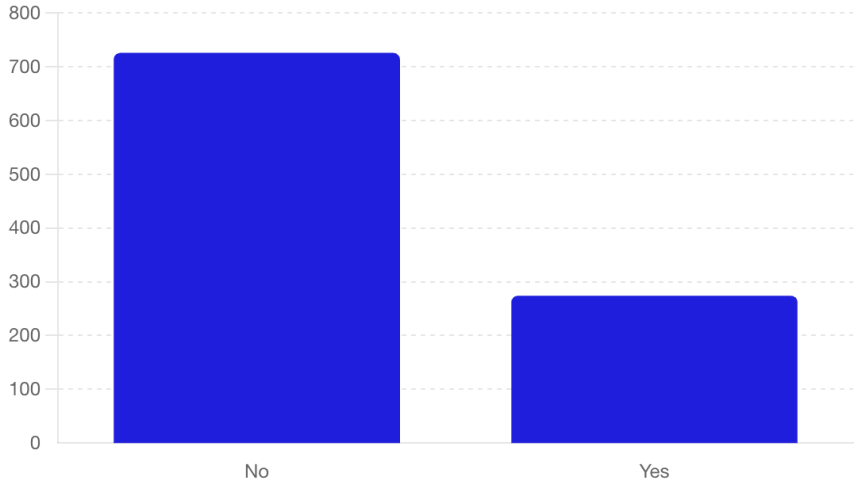
By region



CustCat



By Churn



Sample/Descriptive Characteristics

Here are the results of the sample:

1. **Age**
    - Mean: 41.68 years
    - Standard Deviation: 12.56 years
    - Median: 40 years
    - Range: 18 to 77 years
2. **Years at Current Address**
    - Mean: 11.55 years
    - Standard Deviation: 10.09 years
    - Median: 9 years
    - Range: 0 to 55 years
3. **Gender**
    - Male: 48.3%
    - Female: 51.7%
4. **Level of Education**
    - Levels: 1 to 5 (1 being the lowest and 5 the highest)
    - Mean: 2.67
    - Standard Deviation: 1.22
5. **Income**
    - Mean: $77.54k
    - Standard Deviation: $107.04k
    - Median: $47k
    - Range: $9k to $1668k
6. **Marital Status**
    - Single: 50.5%
    - Married: 49.5%
7. **Region**
    - Region 1: 35.4%
    - Region 2: 32.9%
    - Region 3: 31.7%
8. **Custcat (Customer Category)**
    - Categories: 1 to 4
    - Mean: 2.49
    - Standard Deviation: 1.12
9. **Churn**
    - No: 72.6%
    - Yes: 27.4%

3. **Linear Regression Analysis**

The multiple linear regression analysis results are as follows:

- **R-squared**: 0.103, indicating that approximately 10.3% of the variance in churn can be explained by the predictor variables.
- **Adj. R-squared**: 0.096
- **F-statistic**: 16.23
- **P-value (F-statistic)**: 2.55e−202.55e-202.55e−20

4. For the specific part where churn is chosen as the dependent variable and the other six variables (age, years at current address, gender, level of education, marital status, and region) are the predictors, omitting the variable income, the hypotheses would be:

- **Null Hypothesis**: The regression model with the given predictors does not explain a significant proportion of the variance in churn. (age=address=gender=ed=marital=region=0)
- **Alternative Hypothesis**: The regression model with the given predictors explains a significant proportion of the variance in churn. (βage≠0or βaddress≠0 or βgender≠0 or βed≠0 or βmarital≠0 or βregion≠0)

5. Strength and direction of the relationship

- **R-squared**: 0.103
    - This indicates that approximately 10.3% of the variance in churn can be explained by the predictor variables in the model.
- **Adj. R-squared**: 0.096
    - Adjusted R-squared accounts for the number of predictors in the model and provides a more accurate measure of model performance.
- **F-statistic**: 16.23
    - This value, along with its p-value, indicates that the overall model is statistically significant.

6. **Multiple linear regression analysis for income:**

- **R-squared**: 0.147, indicating that approximately 14.7% of the variance in income can be explained by the predictor variables.
- **Adj. R-squared**: 0.141
- **F-statistic**: 24.43 with a p-value of 7.71e−317.71e-317.71e−31, indicating that the overall model is statistically significant.

7. Overall Hypothesis for the Model

For the multiple linear regression analysis where income is the dependent variable and the predictors are age, years at current address, gender, level of education, marital status, region, and custcat, the hypotheses can be stated as follows:

- **Null Hypothesis**: The regression model with the given predictors does not explain a significant proportion of the variance in income. ($\beta age=\beta address=\beta gender=\beta ed=\beta marital=\beta region=\beta custcat=0$
- **Alternative Hypothesis**: The regression model with the given predictors explains a significant proportion of the variance in income. ($\beta age\neq0$) or $\beta gender\neq0$ or $\beta ed\neq0$ or $\beta marital$ or $\beta region\neq0$ or $\beta custcat\neq0$

8. Interpretation of outputs

1. **Age**
    - **Coefficient**: 2.8412
    - **P-value**: 0.000
    - **Interpretation**: There is a significant positive relationship between age and income. For each additional year in age, the income increases by $2,841.20.
2. **Years at Current Address (address)**
    - **Coefficient**: 0.2007
    - **P-value**: 0.629
    - **Interpretation**: There is no significant relationship between years at the current address and income. The small coefficient and high p-value suggest that years at the current address does not have a meaningful impact on income in this model.
3. **Gender**
    - **Coefficient**: 8.7684
    - **P-value**: 0.164
    - **Interpretation**: There is no significant relationship between gender and income. The coefficient and high p-value suggest that gender does not have a meaningful impact on income in this model.
4. **Level of Education (ed)**
    - **Coefficient**: 13.2951
    - **P-value**: 0.000
    - **Interpretation**: There is a significant positive relationship between the level of education and income. For each higher level of education, the income increases by $13,295.10.
5. **Marital Status**
    - **Coefficient**: -9.6459
    - **P-value**: 0.127
    - **Interpretation**: There is no significant relationship between marital status and income. The negative coefficient and high p-value suggest that marital status does not have a meaningful impact on income in this model.

6. **Region**
    - **Coefficient**: 3.1180
    - **P-value**: 0.421
    - **Interpretation**: There is no significant relationship between region and income. The small coefficient and high p-value suggest that region does not have a meaningful impact on income in this model.
7. **Customer Category (custcat)**
    - **Coefficient**: 8.5407
    - **P-value**: 0.003
    - **Interpretation**: There is a significant positive relationship between customer category and income. For each higher category, the income increases by $8,540.70.

Interpretation of Results

Churn Prediction

- **Strength and Direction: The coefficients from the regression model will indicate the strength and direction of the relationship between each predictor variable and churn.**
- **Significance: The p-values will help determine whether each predictor variable has a statistically significant relationship with churn.**

Income Prediction

- **Strength and Direction: The coefficients from the regression model will indicate the strength and direction of the relationship between each predictor variable and income.**
- **Significance: The p-values will help determine whether each predictor variable has a statistically significant relationship with income.**

Summary of Results

The multiple linear regression analyses conducted to predict both churn and income provide valuable insights into the factors that influence these outcomes for the telecommunications company's customers were informative.

Churn Prediction

In the churn prediction model, the overall R-squared value of 0.103 indicates that approximately 10.3% of the variance in churn can be explained by the predictor variables. The model is statistically significant, as evidenced by an F-statistic of 16.23.

Among the predictors, age, years at current address, and level of education emerged as significant factors. Age and years at the current address both have a significant negative relationship with churn, with coefficients of -0.0051 and -0.0056 respectively, and p-values less than 0.01. This suggests that older customers and those who have lived at their current address for longer periods are less likely to “churn”. Conversely, the level of education has a significant positive relationship with churn (coefficient: 0.0606, p-value: 0.000), indicating that customers with higher levels of education are more likely to churn. Surprisingly, gender, marital status, region, and customer category did not show a significant impact on churn.

Income Prediction

In the income prediction model, the overall R-squared value is 0.147, meaning that approximately 14.7% of the variance in income can be explained by the predictor variables. The model is statistically significant with an F-statistic of 24.43 and a p-value of 7.71. Significant predictors of income include age, level of education, and customer category. Age has a positive relationship with income, with a coefficient of 2.8412 and a p-value of 0.000, indicating that older customers tend to have higher incomes. The level of education also positively correlates with income (coefficient: 13.2951, p-value: 0.000), suggesting that higher educational attainment is associated with higher income levels. Customer category is another significant predictor, with a coefficient of 8.5407 and a p-value of 0.003, implying that higher customer categories are linked to higher incomes. In contrast, years at the current address, gender, marital status, and region were not significant predictors of income, which was unexpected as stability in residence and demographic factors are often considered influential.

Unexpected Findings

There were some surprising results in both models. In the churn prediction model, the lack of significant impact from gender, marital status, and region was unexpected, as these factors are typically thought to influence customer behavior. Similarly, in the income prediction model, the non-significant relationship between years at the current address and income was surprising, as stability in residence is often associated with higher income levels. Additionally, the insignificance of marital status and region in predicting income was also unexpected.

Hypotheses Support and Relationships

The statistical results support the alternative hypotheses for the significant predictors in both models, affirming that age, years at the current address, and education level significantly influence churn, while age, education level, and customer category significantly influence income. The direction of the relationships is consistent with

expectations: older age and longer residence reduce churn, higher education increases churn, and higher age, education, and customer category increase income.

Overall, these analyses provide actionable insights that can help the telecommunications company target efforts to reduce churn and better understand the income levels of their customers.

References:

Elliott, A. C., & Woodward, W. A. (2019). *Mastering SAS for Data Analytics*. Wiley.

UCLA Institute for Digital Research & Education. (n.d.). Regression Analysis. Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-is-multiple-regression/

PennState Eberly College of Science. (n.d.). Multiple Linear Regression Analysis. Retrieved from https://online.stat.psu.edu/stat501/lesson/11