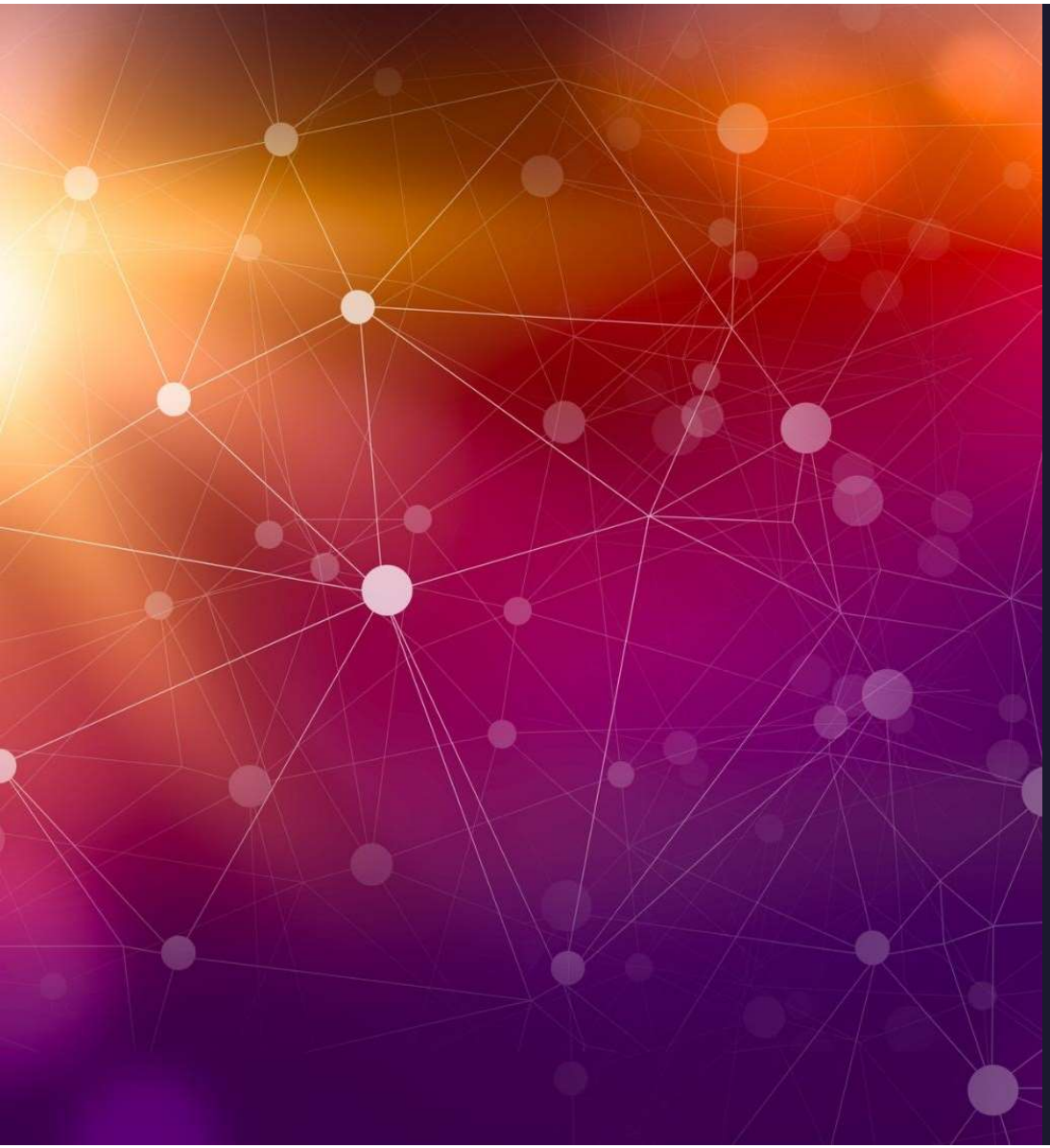


# Predicting Age with Multiple Linear Regression

# Agenda

- Introduction
- Hypothesis
- Data-Analysis Process
- Data Visualizations
- Findings
- Limitations
- Summary

# Introduction

- Kim Fowler
- Completing a Master's degree in Data Analytics from WGU
- Database Developer with 10 years of experience
- Predicting a person's age (years) from a set of health and lifestyle variables

# Hypothesis

# A defined set heath and lifestyle variables statistically significantly affect Age (years).

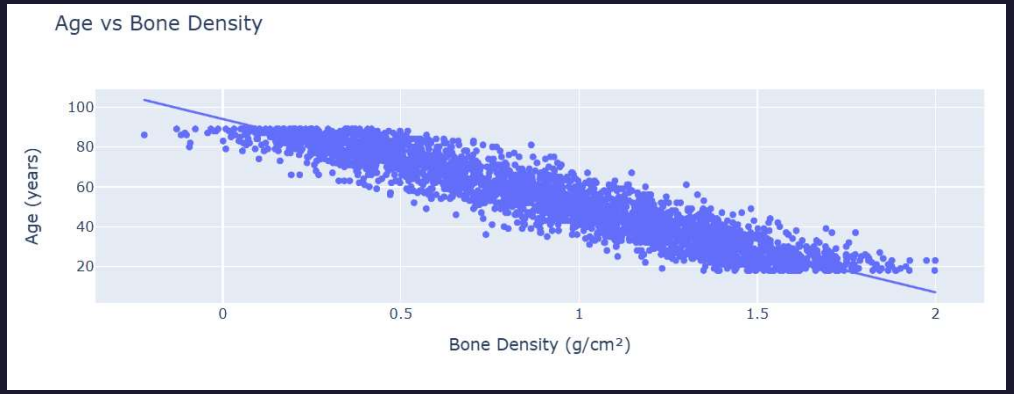
- Hypothesis - The variables Cholesterol Level, Bone Density, Blood Glucose Level, Vision Sharpness, Hearing Ability, and Smoking Status do impact Age (years).
- Alterative Hypothesis - The variables Cholesterol Level, Bone Density, Blood Glucose Level, Vision Sharpness, Hearing Ability, and Smoking Status do not impact Age (years).

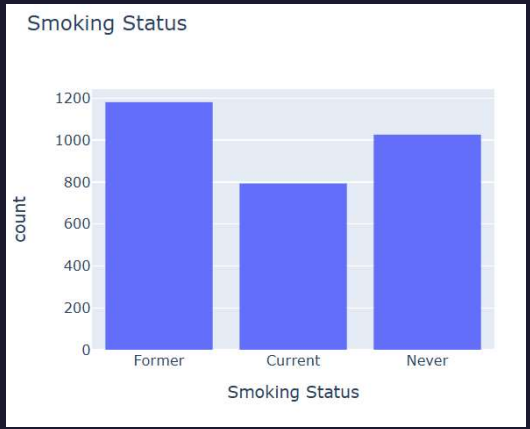
# Data-Analysis Process

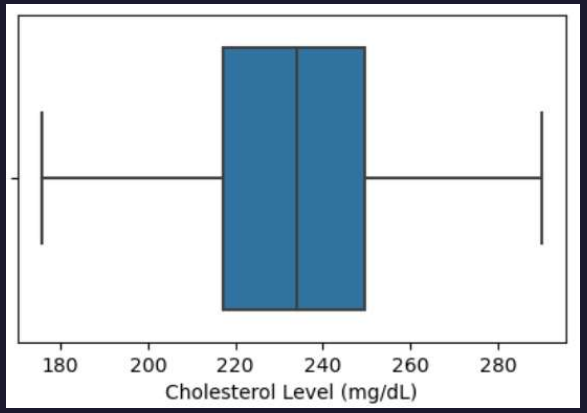
- Data Collection
- Exploratory Data Analysis
- Prepare the Data
- Data Wrangling
- Predictive Modeling
- Feature Selection
- Model Evaluation


Data
Visualizations
Gaining insight into the data

# Visualizations to explore the Data



# Research Findings

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Age (years) | **R-squared:** | 0.928 |
| **Model:** | OLS | **Adj. R-squared:** | 0.928 |
| **Method:** | Least Squares | **F-statistic:** | 6416. |
| **Date:** | Wed, 22 Jan 2025 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:23:27 | **Log-Likelihood:** | -9384.0 |
| **No. Observations:** | 3000 | **AIC:** | 1.878e+04 |
| **Df Residuals:** | 2993 | **BIC:** | 1.882e+04 |
| **Df Model:** | 6 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 73.0186 | 1.620 | 45.084 | 0.000 | 69.843 | 76.194 |
| **Cholesterol Level (mg/dL)** | 0.0335 | 0.005 | 6.876 | 0.000 | 0.024 | 0.043 |
| **Blood Glucose Level (mg/dL)** | 0.0423 | 0.006 | 6.785 | 0.000 | 0.030 | 0.054 |
| **Bone Density (g/cm$^2$)** | -25.8735 | 0.454 | -56.955 | 0.000 | -26.764 | -24.983 |
| **Smoking Status_Never** | -0.5731 | 0.217 | -2.646 | 0.008 | -0.998 | -0.148 |
| **Vision Sharpness** | -32.6142 | 0.929 | -35.093 | 0.000 | -34.436 | -30.792 |
| **Hearing Ability (dB)** | 0.1505 | 0.010 | 15.438 | 0.000 | 0.131 | 0.170 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 0.335 | **Durbin-Watson:** | 2.034 |
| **Prob(Omnibus):** | 0.846 | **Jarque-Bera (JB):** | 0.384 |
| **Skew:** | 0.018 | **Prob(JB):** | 0.825 |
| **Kurtosis:** | 2.957 | **Cond. No.** | 4.41e+03 |

# This model can be used to predict a person's age, thus helping to determine if they are a longevity risk

Formula:

Age (years) = 73.02 + 0.03 (Cholesterol Level) + 0.04 (Blood Glucose Level) - 25.87 (Bone Density)

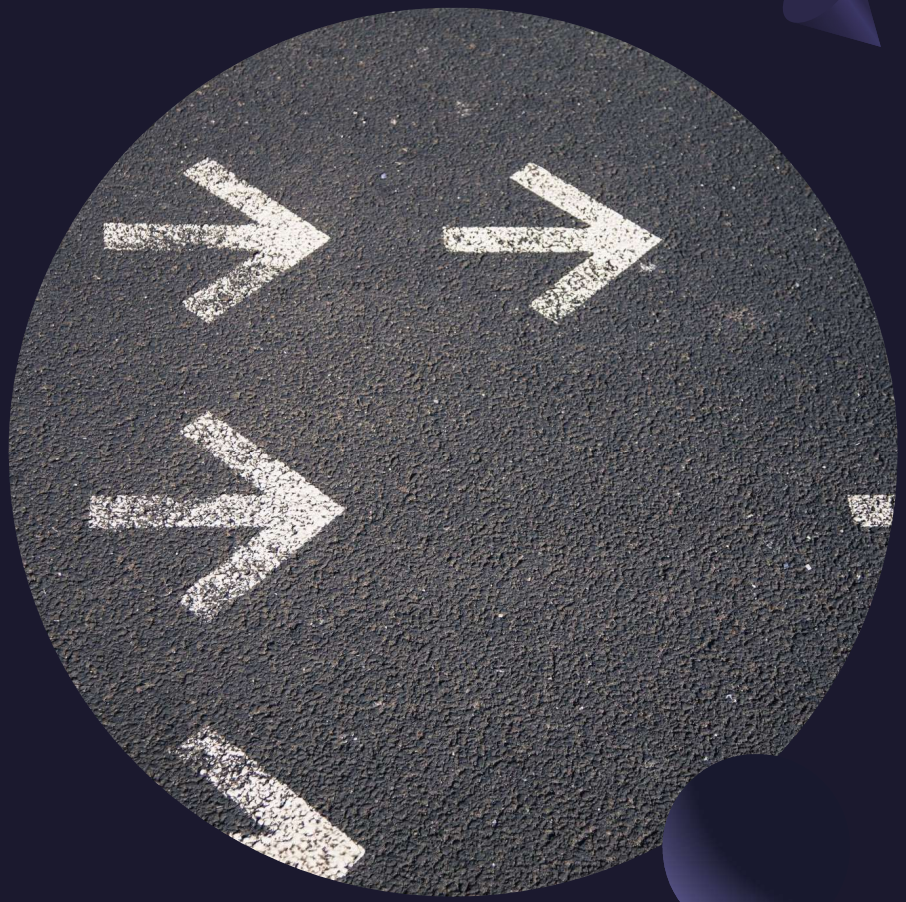
– 32.61 (Vision Sharpness) + 0.15 (Hearing Ability) - 0.57 (Smoking Status)

# Limitations

- Multiple Linear Regression Limitations – sensitive to outliers. Outliers must be handled carefully before performing analysis.
- Data Limitations – The major life insurance company's data may not include all the relevant health and lifestyle variables.

# Next Steps

- Gather comprehensive health and lifestyle information on current customers and their beneficiaries.
- Collect similar health and lifestyle information for new customers and their beneficiaries.



# Conclusion

By using a predictive multiple linear regression model, insights can be gained into whether an individual may be considered a longevity risk.

Identifying individuals who are likely to live longer than expected allows life insurance companies to take proactive measures to mitigate unexpected cash flow issues.