**Data Analysis and Visualization 2 – Project**

**Group Members:** Gurkaranjit, Dhruv, Matthew

**Task 4:**

**1. Discussion of Results from Each Section**

*Task 1:*

We began with an uncleaned dataset from WHO Life Expectancy which contained global records from 2000 to 2015. There was inconsistencies within the data set such as:

* Extra spaces in column names (e.g, “ BMI “ to “BMI”)
* Many missing values (Alcohol\_consumption, schooling and GDP)
* Outliers in columns like Measles

To address these issues, we:

* Created standardized names for the clumns
* Dropped rows that had critical missing data or imputed some using mean/mode
* Added a categorical variable such as region and binary flags for economic status
* Identified extreme outliers and handled them accordingly

The cleaned dataset provided us with a solid foundation to work with which helped with the accuracy of our statistical analysis. Through our visualizations, it showed that:

* Life expectancy is right-skewed, most countries falling between 65 – 80 years
* Developed countries generally had a higher life expectancy with a mean of 80 years
* Developing countries had lower with a mean of 69 years.
* North America and EU led the regions in life expectancy while Africa lagged behind the rest significantly

The correlation heatmap revealed to us that there was a strong linear relationship between the variables. The most notable ones were:

* Life\_expectancy had a strong negative correlation with Adult\_mortality (-0.95) and Infant\_deaths (-0.92).
* There were positive relationships shown with schooling (+0.73), BMI (+0.60) and GDP (+0.58)

*Task 2:*

We proposed two research questions which were:

1. Is there a significant difference in the life expectancy between developed and developing countries?
2. Has adult mortality changed significantly between 2000 and 2015?

In order to test these questions, we:

* Used a two sample t-test for the first question and found that there was a highly significant difference (p < 2.2e-16). Developed countries had a mean life expectancy of ~78.5 years vs ~66.3 years in developing ones
* For the second question, we used a paired t-test and found that there was a significant improvement in adult mortality from 2000 to 2015. The mean difference was approximately 54 deaths per 1000, p < 0.001

These results aligned well with global trends on improving health outcomes. It also highlights the disparities between economic classifications.

*Task 3:*

When creating our model, we used multiple linear regression to predict Life\_expectancy based on:

* Alcohol\_consumption, Schooling, GDP, BMI, Adult\_mortality and Infant\_deaths

This model results came back very strong:

* Adjusted R^2 = 0.977, which means that 97.7% of the variation in life expectancy was explained
* All predictors were also statistically significant (p < 0.001)
  + Every extra year of schooling resulted in an increase of life expectancy by ~0.16 years
  + Every increase of 1000 in GDP added ~0.027 years
  + Each infant death per 1000 reduced the life expectancy by ~0.13 years

**2. Limitations of Our Analysis**

Even though our analysis was proven strong, there still existed some limitations:

* Missing values in original data: there were missing values or values we had to drop which could reduce the accuracy in our analysis
* GDP skew and inflation: GDP was used as provided in the dataset. In reality, we could have to compare GDP over multiple years and adjust inflation accordingly
* Regional Imbalance: There were some regions that contained more data than others which could possibly skew regional comparisons
* Binary classification of economic status: The countries were only labeled as developed or developing. We could not account for nuances like emerging economies or economic transitions

**3. Suggestions for Improvement**

To improve our analysis further, we could have:

* Use more recent data (post 2015) for more relevancy
* Use normalized predictors like GDP to reduce skewness
* Expanding the classification of countries into more than two groups. This would make for better economic stratification
* We could have also incorporated a machine learning model to explore more complex relationships
* A better model representation between GDP and Life\_expectancy due to the non-linear relationship in the plot from task 3. This suggests a polynomial regression might be needed

**Final Conclusion**

Through our analysis, we successfully identified and confirmed key factors related to socioeconomic and health which influenced life expectancy across the globe. We also confirmed that there was a large disparity between developed and developing countries, significant improvement in adult mortality over time, and education with GDP were powerful predictors for national longevity. These findings provide critical insights aimed at reducing global health inequality.