**Executive Summary**

The summary of the findings in the portfolio of my statistical analyses and the reflections on how I demonstrated all the learning outcomes is presented in this section.

I applied relevant statistical techniques on 3 different datasets. The datasets include Wine Quality Dataset obtained from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/>); Melbourne Housing Dataset obtained from Kaggle (<https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market>), and HR Employee Attrition Dataset obtained from Kaggle (<https://www.kaggle.com/code/evaaasong/ibm-hr-analytics-employee-attrition-performance/data>).

From the Wine Quality Dataset, I was able to develop the hypothesis below:

H01: There is no significant relationship between the quality of wine and its attributes.

HA1: There is a significant relationship between the quality of wine and its attributes.

A few of the assumptions made about this data include:

1. There is absence of multicollinearity in the data. This means there won’t be situations in which the predictors have very high correlations among them, and the independent variables would not be explaining each other. The presence of multicollinearity in data can lead to a spurious regression result.
2. There is linearity between the dependent variables and independent variables.

The assumption of no multicollinearity was first checked with pairplots made with the ggpairs function from the GGally R package. It was found that no correlation coefficient was higher than 0.8, an indication that there is no multicollinearity in the data. After the multiple linear regression was performed, a formal multicollinearity test was performed on the model with the vif function. It was used because it provides a better way of checking for multicollinearity rather than merely looking at the correlation coefficients. It was found that none of the independent variables has a vif higher than 10. This further confirmed that there is no multicollinearity in the data.

The assumption of linearity was tested using the Q-Q plot generated as part of the output for diagnostic test on the model. It was found that most of the points do not lie far away from the line, so the data comes close to a normally distributed one.

To enable me test the hypothesis, I conducted a multiple linear regression, with quality as the dependent variable, and attributes such as fixed acidity, citric acid, chlorides, pH, density, alcohol, etc. The outcome of the hypothesis tested using the Wine Quality Dataset revealed that there is a significant relationship between quality of wine and its attributes such as volatile acidity (p-value = 0.000), chlorides (p-value = 0.001), total sulfur dioxide (p-value = 0.001), and alcohol (p-value = 0.000). The p-value of the F-test generated with the regression output was 0.000, a value less than 0.05. This showed that overall, there is a statistically significant relationship between the quality of wine and its attributes, we therefore reject the null hypothesis that there is no significant relationship between quality of wine and its attributes.

From the Melbourne Dataset, I was able to develop the hypothesis below:

H02: There is no significant relationship between the price of a house and its characteristics.

HA2: There is a significant relationship between the price of a house and its characteristics.

Like the Wine Quality Dataset, the assumptions made about the dataset include:

1. There is absence of multicollinearity in the data.
2. There is linearity between the dependent variables and independent variables.

To check the assumption of no multicollinearity, pairplots was also made for this data. It was found that no correlation coefficient was higher than 0.8, which shows there is suspicion of serious multicollinearity in the data.

After the multiple linear regression was performed, a formal multicollinearity test was also performed on the model with the vif function. This is to ensure that the assumption of no multicollinearity truly holds. None of the variables had a vif exceeding 10, indicating that multicollinearity is absent in the data.

The assumption of linearity in the regression model was tested using the Q-Q plot generated as part of the output for diagnostic test on the model. It was found that most of the points do not lie far away from the line, so the data comes close to a normally distributed one.

The outcome of the hypothesis tested using the Melbourne Housing Dataset revealed that there is a significant relationship between the price of a house and its characteristics such as Rooms (p-value = 0.000), Distance (p-value = 0.001), Bedroom (p-value = 0.001), Bathroom (p-value = 0.000), Car (0.000), and BuildingArea (p-value = 0.001). The p-value of the F-test generated with the regression output for the Melbourne Housing Dataset was 0.000, a value less than 0.05. This showed that overall, there is a statistically significant relationship between the price of a house and its characteristics. I therefore rejected the null hypothesis that there is no significant relationship between the price of a house and its characteristics.

The Melbourne Housing dataset contains lots of missing values. A few of the variables also contained unwanted characters. To fill the rows with missing values in the data, median imputation was performed on each of the affected columns. There are many ways imputation can be performed. But median imputation was used because it helps to prevent the results of the analysis from being skewed. The unwanted characters were also removed from the affected after which they were properly converted to a numeric variable, a form suitable for the analysis conducted.

I was interested in understanding the impact attributes such as age of employees, education, distance from their homes, monthly income, and so on have on their attrition, so I chose the HR Attrition Dataset with the aim of conducting a regression analysis to for this purpose.

The hypothesis developed was:

H03: Factors such as age, distance from home, monthly income, do not have an impact on attrition.

HA3: Factors such as age, distance from home, monthly income, do have an impact on attrition.

Since attrition is a categorical variable with two values: “Yes” and “No”, logistic regression was employed for the purpose of this analysis.

It was found that factors such as age (p-value =0.02), distance from home (p-value = 0.002), environmental satisfaction (p-value = 0.000), and monthly income (0.03) have impact on attrition. Therefore, the null hypothesis that factors such as age, distance from home, monthly income, do not have an impact on attrition was rejected.

Furthermore, I was interested in testing whether the monthly income earned by male employees differ significantly from the monthly income earned by female employees, and whether the age of employees who attrit is significantly different from the age of employees who do not attrit. So, I developed these hypotheses below:

H04: There is no significant difference in the monthly income of employees across gender.

HA4: There is a significant difference in the monthly income of employees across gender.

H05: There is no significant difference in the age of the employees across different attrition status.

HA5: There is a significant difference in the age of the employees across different attrition status.

Independent samples T-test was the suitable test for the latest two hypotheses stated above. For the first of these two hypotheses, the p-value obtained was 0.22, a value greater than 0.05. Therefore, I rejected the alternative hypothesis that there is a significant difference in the monthly income of employees across gender. What this means is that average monthly income does not vary much between the two genders.

For the second of these two hypotheses, the p-value obtained was 0.00, a value less than 0.05. Therefore, I rejected the null hypothesis there is a significant difference in the age of the employees across different attrition status. What this means is that the mean age of employees who attrit is statistically different from those who do not attrit.

**Reflection**

By conducting this analysis, I was able to achieve the learning outcome set for this assessment. I obtain a comprehensive understanding of different statistical concepts such as p-values, t-test, multicollinearity, logistic regression, etc. It excites me to know that these statistical concepts have wider applications in our everyday life. Statistical concepts are not limited to a specific field. They have an application in business, marketing, engineering, human resources, and so on. In this assessment for instance, I was able to apply statistical concept to the business world. I could see various factors that contribute to attrition among workers of different companies. I was also able to understand that different attributes can determine the price at which a house can be sold, as the in the case of Melbourne Housing.

Developing hypotheses can help provide direction to an analysis. With the right research questions, I was able to develop useful hypotheses which also inform the decision of the statistical tests I applied on the data selected for this task. Problems that seem complex were broken into bits and solved with the appropriate statistical, and the results properly communicated.