**Analysing Biological Signals to Predict Smoking Behaviour**

**Introduction**

In England, smoking is the leading factor behind avoidable health issues and untimely death, contributing to approximately 74,000 deaths annually. Smoking significantly lowers life expectancy and damages almost all of the body's organs. Smoking causes multiple tumours in various organs such as the throat, mouth, kidney, bladder, liver, and cervix, in addition to lung cancer, respiratory disorders, heart disease, and other diseases. Around 8 million people worldwide are estimated to die every year from tobacco usage (including second-hand smoke exposure). There's an estimation that by the end of the 21st century, tobacco use will have resulted in the deaths of one billion individuals. For every smoking-related death, roughly 30 smokers are afflicted by diseases caused by smoking. An estimated 506,100 NHS hospitalizations among people 35 or older in England were caused by smoking in 2019–20. Smoking is estimated to generate an annual expense of £1.9 billion for the NHS in England **(ASH, 2023)**.

According to data from the U.S. National Health Survey, people who give up smoking, regardless of age, have a lower risk of dying from an illness related to smoking than those who continue to smoke. Individuals who quit smoking before reaching the age of 40 can decrease their likelihood of early death due to smoking-related diseases by approximately 90%. For those who cease smoking between the ages of 45 and 54, the risk of premature death is reduced by nearly two-thirds. When compared with people who continue to smoke, people who give up smoking, regardless of age, have significantly longer life expectancies **(NIH, 2013)**.

**The Objective**

Our report's objective is to analyse body signals to identify smokers and gain insights. We'll achieve this by using the best statistical methods tailored to our dataset.

The main purpose of accurately distinguishing between smokers and non-smokers is to better understand the link between smoking and various health issues. Smoking is a significant risk factor for diseases like lung cancer, heart problems, and breathing issues. When we correctly identify smokers and non-smokers, it enables researchers to study the impact of smoking on health more precisely. This, in turn, helps in creating effective public health initiatives to reduce smoking rates and prevent health problems.

This classification also has practical consequences. For example, insurance companies may charge higher premiums to smokers due to their greater health risks, and employers might choose not to hire smokers because of productivity concerns. To accomplish our goal, we will employ strong statistical methods to ensure accurate identification of smokers and non-smokers **L&C. (2023)**.

**Dataset**

The dataset comprises fundamental biological health signal information obtained from routine health check-up results. The objective is to identify the presence or absence of a smoking habit using these biological signals.

Data shape: (55692, 27)

The variables of the dataset:

The features of the dataset:

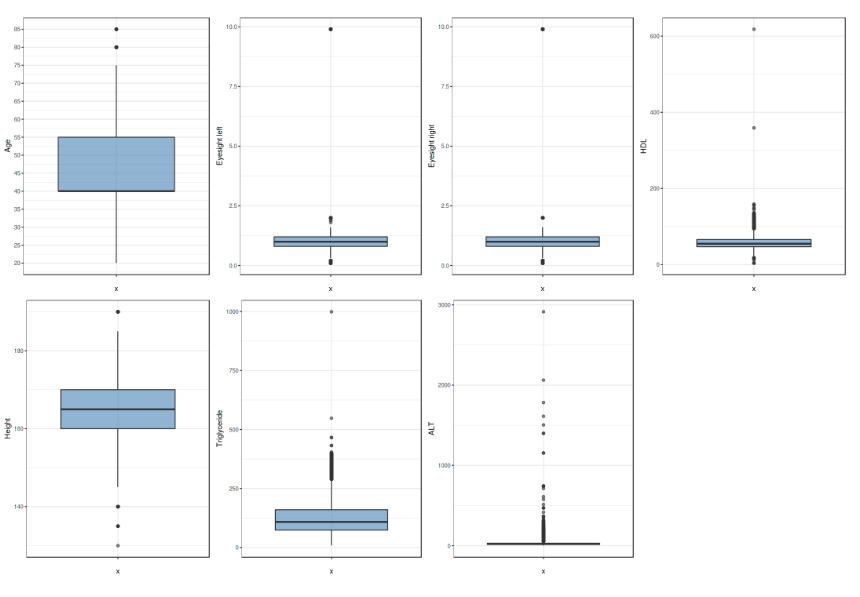
* ID: serial number given to a person.
* gender: the gender of a person being either female (0) or male (1).
* age: Age groups with a 5-year interval.
* height(cm)
* weight(kg)
* waist(cm)
* eyesight(left): Denoted by a value between 0.1-2.5, or 9.9
* eyesight(right): Denoted by a value between 0.1-2.5, or 9.9
* hearing(left): hearing of the person's ear (1 - normal, 2 - abnormal)
* hearing(right): hearing of the person's ear (1 - normal, 2 - abnormal)
* fasting blood sugar: blood sugar before meals level
* systolic: Blood Pressure
* relaxation: Blood Pressure
* triglyceride
* LDL: Type of a cholesterol
* HDL: Type of a cholesterol
* Cholesterol: total
* Haemoglobin
* serum creatinine
* Urine protein
* AST:
* ALT:
* Gtp: γ-GTP
* dental caries
* tartar: tartar status
* oral: Oral Examination status (s whether the examinee accepted the oral examination).
* smoking: smoking status of a person (1 - smoking, 0 - non-smoking)

Our data contains no null values among all 55692 rows. The dataset consists of 26 features (excluding the ID column) with majority of them having numerical values. There are 3 categorical features, namely gender, oral, and tartar.

**Data Cleaning**

Data cleaning is all about finding and fixing errors in your data to make it better. An error is when the data doesn't match what it's supposed to represent. So, in data cleaning, you look at your data, find any mistakes, and correct them. It's like tidying up your dataset, making it nice and clean. People also call it data scrubbing **(Bhandari, 2022)**

We verified for any unusual data points and the columns which have suspected outliers are age, eyesight left and right, HDL, height, triglyceride and ALT



Based on the boxplot summaries, it is evident that the count of outliers for all the suspected variables exceeds 200, indicating that they should not be classified as outliers.

We checked for missing values in our data with the help of python. There is no missing value in our data. We removed some columns like ID, height and waist from our data set.

**Analysis of Data**

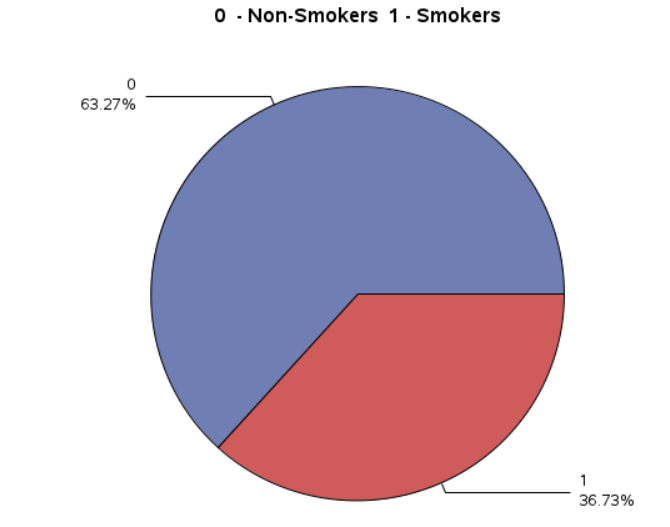
It is essential to visualize the data using plots like bar charts or box plots to gain an initial understanding of the relationship between variables. Data visualization can help guide your choice of statistical analysis and provide valuable insights.

Dataset

* Rows= 55692, Columns = 24

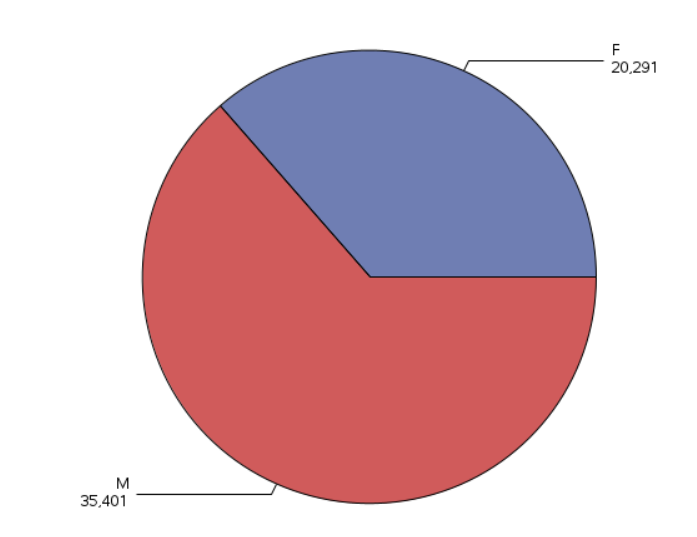
The percentage of smoking status in the dataset:

* Non-smoking = 63.3%
* Smoking =36.7 %

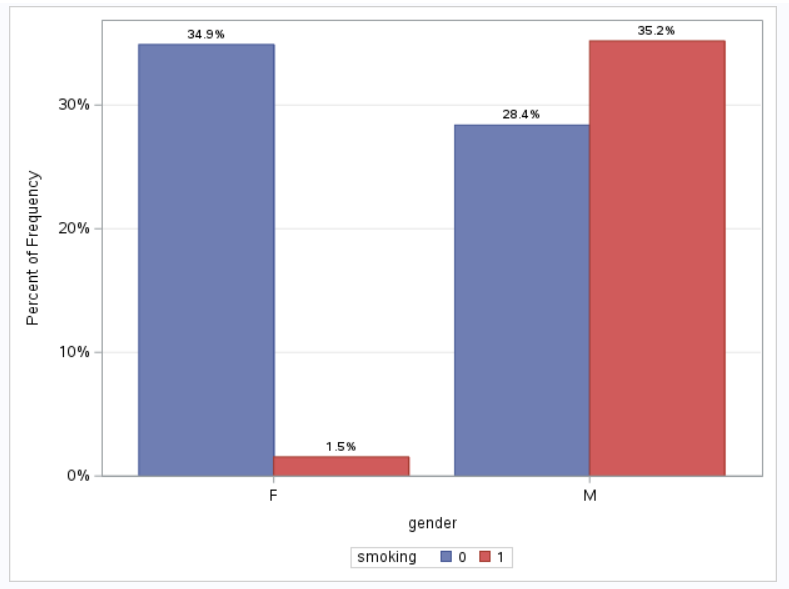


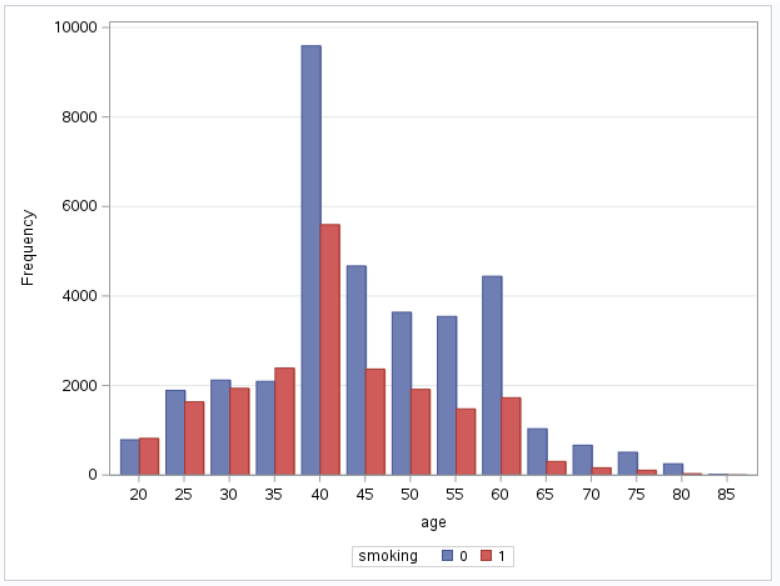
The percentage gender in the dataset:

* Female = 36.4%
* Male =63.6 %



Only 1.5% of the smokers are Females and 35.2% are Males.





The average data

Female

* Average age of Smoking = 46
* Average age of Non-smoking = 49

Male

* Average age of Smoking = 41
* Average age of Non-smoking = 42

About two-thirds of the group are men, and one-third are women. Additionally, the majority of people in the group don't smoke, with only 36% being smokers.

**Case Study 1**

**Problem Statement –** Is there any notable impact of individuals age on their smoking status.

**Solution:**

Null Hypothesis(H0): There is no significant effect of age on smoking status.

Alternative Hypothesis(H1): There is a significant effect of age on smoking status.

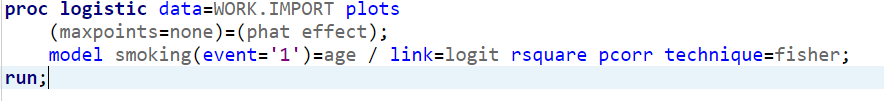
Since age variable is continuous and smoking status is binary (e.g., 0 for non-smoker, 1 for smoker), logistic regression method is appropriate to test the hypothesis.

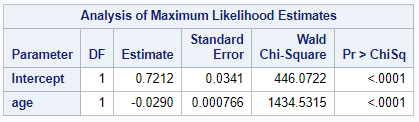
Logistic regression works by estimating the chance of an event happening, true/false or yes/no, based on a set of input data. Because we're dealing with probabilities, the result is always between 0 and 1. In logistic regression, we use a mathematical transformation called the logit, which represents the natural logarithm of the odds **(IBM, 2016)**.

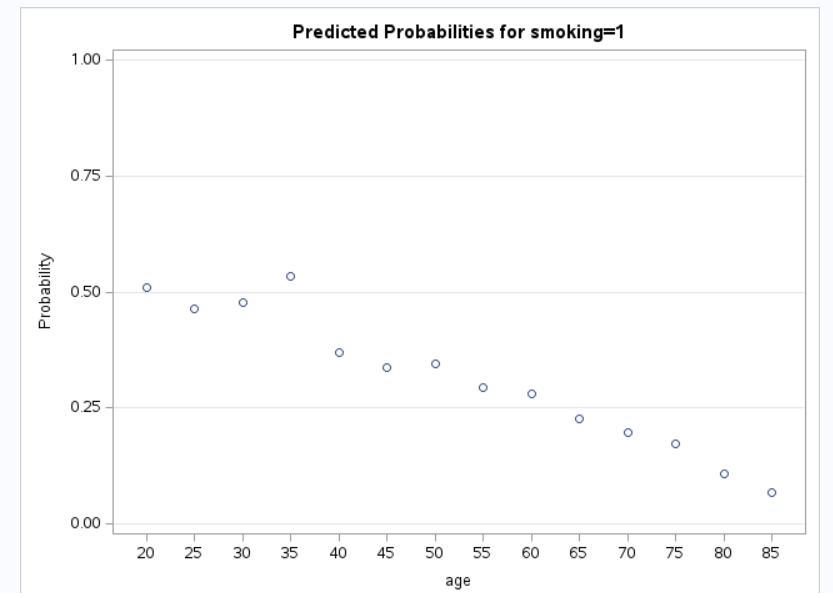
In our problem, logistic regression helps in testing whether age is a significant predictor of smoking status.

We will take the help of SAS tool and import our data in it. Below are the code and the screenshot of the table and graph we generated from the SAS tool.

Code:







From the above table and graph, we make the following analysis.

1. Intercept (0.7212):
   * The intercept signifies the estimated log-odds of being a smoker when age is set to zero. However, this interpretation lacks practical significance, as age cannot realistically be zero.
2. Age (-0.0290):
   * The coefficient for "age" is -0.0290.
   * This negative coefficient indicates that as age increases by one unit (presumably one year), the log-odds of being a smoker decrease by 0.0290 units.

To make the interpretation more intuitive, we can transform the log-odds into odds ratios:

* The odds ratio for "age" can be calculated as exp (-0.0290) ≈ 0.9723.

Interpretation:

* For every one-year increase in age, the odds of being a smoker decrease by approximately 2.77%.

In other words, according to this logistic regression model, as individuals get older, they are less likely to be smokers. This is supported by the negative coefficient and odds ratio for age. The odds ratio less than 1 indicates a decrease in the odds of smoking with increasing age.

The highly significant p-values ("<.0001") for and age coefficient suggest that age have a significant impact on smoking status in this model. Therefore, the alternative hypothesis(H1) is true and we will reject the null Hypothesis(H0).

Additionally in logistic regression, we can use the formula to calculate the predicted log-odds of the outcome variable. Given the coefficients and standard errors provided in our results, we can use the formula:

Log-Odds (logit) = Intercept + (Coefficient \* Age)

Let's use the coefficients and standard errors from your results to calculate the log-odds for a given age.

In our results:

* Intercept: 0.7212
* Age Coefficient: -0.0290

Suppose we want to calculate the log-odds of being a smoker for a person of age 40. We would use the formula as follows:

Log-Odds (logit) = 0.7212 + (-0.0290 \* 40) = 0.7212 - 1.16 = -0.4388

The log-odds for a 40-year-old individual being a smoker is approximately -0.4388.

To convert the log-odds to probabilities, we can use the logistic function (also known as the sigmoid function):

Probability (Smoker) = 1 / (1 + e^(-logit))

Where "e" corresponds to the mathematical constant known as the base of the natural logarithm, approximately equivalent to 2.71828.

Probability (Smoker) = 1 / (1 + e^(-(-0.4388)) = 1 / (1 + e^(0.4388)) ≈ 0.3938

So, the estimated probability of a 40-year-old individual being a smoker is approximately 0.3938, or about 39.38%.

We can use this formula to calculate the log-odds and probabilities for different ages based on the coefficients and standard errors provided in your logistic regression results. This allows us to make predictions about the likelihood of being a smoker for individuals at different ages.

**Top of Form**

**Case Study 2**

**Problem Statement -** Is there any significant impact of smoking on various body signals like blood pressure, cholesterol, triglyceride, dental and liver.

**Solution:**

We will conduct a step-by-step comparison of smoking status with various body signals. For systolic blood pressure, we will provide a detailed t-test analysis, while for the rest of the variables, we will present the t-test results without reiterating the entire methodology.

**Smoking and Blood Pressure**

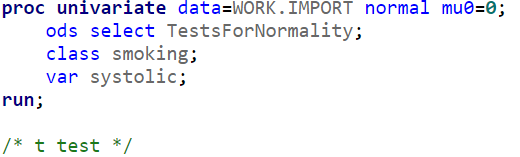
T-test is a statistical analysis employed for making inferences about whether there is a significant difference between the means of two groups and how they are related. **(Hayes, 2023)**.

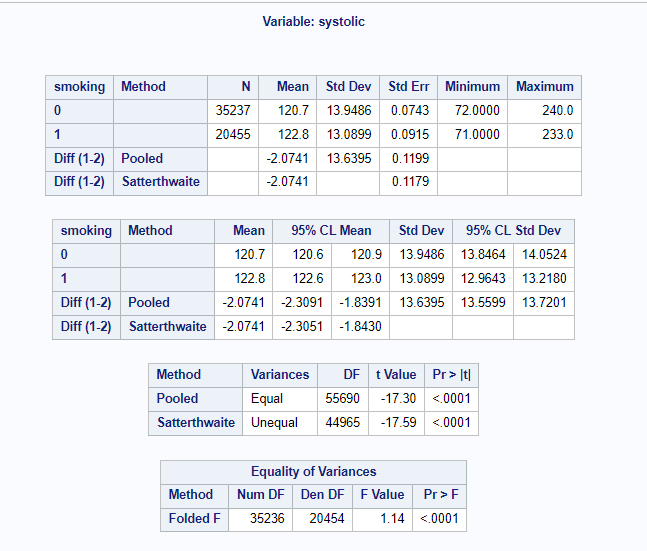
Here are some reasons why a t-test is a good choice for our problem of examining the impact of smoking on blood pressure:

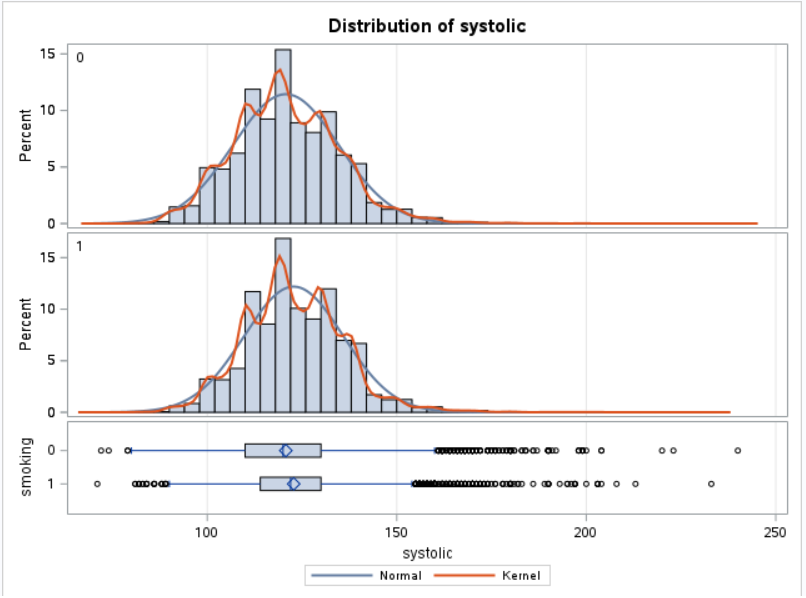
1. Two Groups Comparison: In our problem, we have two distinct groups: smokers and non-smokers. A t-test is designed specifically for comparing means between two groups, making it an appropriate choice.
2. Continuous Outcome Variable: Blood pressure is a continuous variable that is typically normally distributed. The t-test assumes that the data follow a normal distribution, which is often the case with continuous physiological measures like blood pressure.

Null Hypothesis (H0): The mean systolic blood pressure does not vary between smokers and non-smokers.

Alternative Hypothesis (H1): There exists a notable distinction in the mean systolic blood pressure between smokers and non-smokers.







According to the result presented above, we can conclude that our variances are not equal because the p-value in the equality of variance table is less than 0.05.

So, let’s interpret the result according to the unequal variance result.

1. t-Value: The t-value of -17.59 is a measure of how different the means of the two groups are relative to the variation within each group. In this case, a very negative t-value indicates that there is a significant difference in systolic BP between the two groups.
2. p-Value (Pr > |t|): The p-value is less than 0.05, which indicates that there is a highly significant difference between the mean of systolic BP for smokers and non-smokers.
3. Means and Confidence Intervals: The means for the two groups are as follows:
   * For non-smokers (smoking status 0): Mean systolic BP = 120.7
   * For smokers (smoking status 1): Mean systolic BP = 122.8

These means suggest that, on average, smokers have slightly higher systolic BP than non-smokers.

In brief, according to the t-test outcomes, there is compelling evidence to indicate that there is a substantial difference in systolic blood pressure associated with smoking status. Smokers, on average, have slightly higher systolic BP compared to non-smokers, with the difference estimated to be around 2 units.

**Smoking and other body signals like cholesterol, triglyceride, dental and liver.**

According to the results of the t-test, we have determined that the null hypothesis is rejected for all body signals. This indicates a significant difference in the mean values of body signals between smokers and non-smokers.

**Conclusion**

In conclusion, our analysis reveals that smoking is linked to significant health risks:

1. Age and Smoking: People tend to smoke more when they are younger, and smoking rates decrease after the age of 40.
2. Smoking and High Blood Pressure: Smoking is often associated with high blood pressure, which can lead to heart diseases.
3. Smoking and Cholesterol: Smokers tend to have lower levels of good cholesterol, which is important for heart health.
4. Triglycerides: Smokers have higher triglyceride levels, which can be risky for their health.
5. Liver Function: A marker of liver function called GTP is often at risky levels in smokers, indicating potential liver health issues.

These findings emphasize the importance of addressing smoking as a public health concern and promote quitting smoking to reduce health risks.

**References**

Action on Smoking and Health (ASH). (2023). *Smoking Statistics*. <https://ash.org.uk/uploads/Smoking-Statistics-Fact-Sheet.pdf?v=1697728811>

National Institutes of Health (NIH). (2023). *21st-century hazards of smoking and benefits of cessation in the United States*. <https://pubmed.ncbi.nlm.nih.gov/23343063/>

L&C. (2023). *How does smoking affect life insurance ?*. <https://www.landc.co.uk/mortgage-guides/how-does-smoking-affect-life-insurance/>

Bhandari, (2022). *Data Cleaning | A Guide with Examples & Steps*. Scribbr. <https://www.scribbr.co.uk/research-methods/data-cleaning/>

IBM. (2016). *What is logistic regression?*. IBM. <https://www.ibm.com/topics/logistic-regression>

Hayes, (2023). *T-Test: What It Is With Multiple Formulas and When To Use Them*. investopedia. <https://www.investopedia.com/terms/t/t-test.asp>

Data Set - Kaggle. (2022). *Body signal of smoking*. Kaggle. <https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking/data>

Graph - [SAS. SAS. https://welcome.oda.sas.com/](file:///D:\UK_University\Assignments\Statistics_Assignment\SAS.%20SAS.%20https:\welcome.oda.sas.com\)