

# A Gentle Introduction to Statistics Using SAS® Studio



Ron Cody

The correct bibliographic citation for this manual is as follows: Cody, Ron. 2019. *A Gentle Introduction to Statistics Using SAS® Studio*. Cary, NC: SAS Institute Inc.

### **A Gentle Introduction to Statistics Using SAS® Studio**

Copyright © 2019, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-64295-541-5 (Hardcover)

ISBN 978-1-64295-532-3 (Paperback)

ISBN 978-1-64295-533-0 (Web PDF)

ISBN 978-1-64295-534-7 (EPUB)

ISBN 978-1-64295-535-4 (Kindle)

All Rights Reserved. Produced in the United States of America.

**For a hard copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government License Rights; Restricted Rights:** The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

October 2019

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

# Contents

<b>About This Book .....</b>	<b>vii</b>
<b>About The Author .....</b>	<b>ix</b>
<b>Acknowledgments .....</b>	<b>xi</b>
<b>Chapter 1: Descriptive and Inferential Statistics .....</b>	<b>1</b>
Overview .....	1
Descriptive Statistics .....	1
Inferential Statistics .....	2
Summary of Statistical Terms .....	4
<b>Chapter 2: Study Designs .....</b>	<b>7</b>
Introduction .....	7
Double-Blind, Placebo-Controlled Clinical Trials .....	7
Cohort Studies .....	8
Case-Control Studies .....	9
Conclusion .....	9
<b>Chapter 3: What Is SAS University Edition? .....</b>	<b>11</b>
Introduction .....	11
How to Download SAS University Edition .....	12
Conclusion .....	20
<b>Chapter 4: SAS Studio Tasks .....</b>	<b>21</b>
Introduction .....	21
Using the Built-in Tasks .....	23
Taking a Tour of the Navigation Pane .....	24
Exploring the LIBRARIES Tab .....	25
Conclusion .....	29
<b>Chapter 5: Importing Data into SAS .....</b>	<b>31</b>
Introduction .....	31
Exploring the Utilities Tab .....	32
Importing Data from an Excel Workbook .....	33
Listing the SAS Data Set .....	38
Importing an Excel Workbook with Invalid SAS Variable Names .....	39

Importing an Excel Workbook That Does Not Have Column Headings .....	40
Importing Data from a CSV File.....	41
Shared Folders (Accessing Data from Anywhere on Your Hard Drive) .....	42
Conclusion.....	42
<b>Chapter 6: Descriptive Statistics – Univariate Analysis.....</b>	<b>43</b>
Introduction.....	43
Generating Descriptive Statistics for Continuous Variables.....	44
Investigating the Distribution of Horsepower .....	49
Adding a Classification Variable in the Summary Statistics Tab.....	52
Creating a Filter Within a Task .....	54
Creating a Box Plot .....	57
Conclusion.....	59
Chapter 6 Exercises .....	59
<b>Chapter 7: One-Sample Tests .....</b>	<b>61</b>
Introduction.....	61
Getting an Intuitive Feel for a One-Sample t Test .....	61
Performing a One-Sample t Test.....	62
Nonparametric One-Sample Tests .....	69
Conclusion.....	71
Chapter 7 Exercises .....	72
<b>Chapter 8: Two-Sample Tests .....</b>	<b>73</b>
Introduction.....	73
Getting an Intuitive Feel for a Two-Way t Test .....	73
Unpaired t Test (t Test for Independent Groups).....	74
Describing a Two-Sample t Test .....	75
Nonparametric Two-Sample Tests .....	82
Paired t Test.....	86
Conclusion.....	90
Chapter 8 Exercises .....	90
<b>Chapter 9: Comparing More Than Two Means (ANOVA) .....</b>	<b>93</b>
Introduction.....	93
Getting an Intuitive Feel for a One-Way ANOVA.....	94
Performing a One-Way Analysis of Variance .....	94
Performing More Diagnostic Plots .....	102
Performing a Nonparametric One-Way Test .....	104
Conclusion.....	109
Chapter 9 Exercises .....	109
<b>Chapter 10: N-Way ANOVA .....</b>	<b>111</b>
Introduction.....	111
Performing a Two-Way Analysis of Variance .....	112
Reviewing the Diagnostic Plots .....	117

Interpreting Models with Significant Interactions .....	120
Investigating the Interaction .....	123
Conclusion .....	124
Chapter 10 Exercises .....	124
<b>Chapter 11: Correlation.....</b>	<b>127</b>
Introduction .....	127
Using the Statistics Correlation Task.....	127
Generating Correlation and Scatter Plot Matrices .....	130
Correlations among Variables in the Fish Data Set.....	134
Interpreting Correlation Coefficients.....	136
Generating Spearman Non-Parametric Correlations.....	137
Conclusion .....	138
Chapter 11 Exercises .....	139
<b>Chapter 12: Simple and Multiple Regression .....</b>	<b>141</b>
Introduction .....	141
Getting an Intuitive Feel for Regression .....	142
Describing Simple Linear Regression .....	143
Understanding How the F Value Is Computed .....	148
Investigating the Distribution of the Residuals.....	149
Measures of Influence .....	150
Demonstrating Multiple Regression.....	151
Running a Simple Linear Regression Model with Endurance and Pushups.....	152
Demonstrating the Effect of Multi-Collinearity .....	154
Demonstrating Selection Methods .....	157
Using a Categorical Variable as a Predictor in Model .....	160
Conclusion .....	161
Chapter 12 Exercises .....	161
<b>Chapter 13: Binary Logistic Regression .....</b>	<b>163</b>
Introduction .....	163
Describing the Risk Data Set .....	163
Running a Binary Logistic Regression Model with a Single Predictor Variable .....	164
A Discussion about Odds Ratios.....	168
Editing SAS Studio-Generated Code .....	170
Using a Continuous Variable as a Predictor in a Logistic Model .....	171
Running a Model with Three Classification Variables.....	172
Conclusion .....	174
Chapter 13 Exercises .....	175

<b>Chapter 14: Analyzing Categorical Data .....</b>	<b>177</b>
Introduction.....	177
Describing the Salary Data Set .....	177
Computing One-Way Frequencies .....	178
Creating Formats.....	181
Producing One-Way Tables with Formats .....	183
Reviewing Relative Risk, Odds Ratios, and Study Designs .....	184
Creating Two-Way Tables .....	185
Using Formats to Reorder the Rows and Columns of a Table .....	189
Computing Chi-Square from Frequency Data .....	192
Analyzing Tables with Low Expected Values .....	194
Conclusion.....	196
Chapter 14 Exercises .....	196
<b>Chapter 15: Computing Power and Sample Size .....</b>	<b>199</b>
Introduction.....	199
Computing Sample Size for a t Test.....	199
Calculating the Sample Size for a Test of Proportions .....	204
Computing Sample Size for a One-Way ANOVA Design .....	208
Conclusion.....	210
Chapter 15 Exercises .....	211
<b>Odd-Numbered Exercise Solutions.....</b>	<b>213</b>
Chapter 6 Solutions.....	213
Chapter 7 Solutions.....	214
Chapter 8 Solutions.....	215
Chapter 9 Solutions.....	216
Chapter 10 Solutions.....	217
Chapter 11 Solutions.....	218
Chapter 12 Solutions.....	219
Chapter 13 Solutions.....	220
Chapter 14 Solutions.....	221
Chapter 15 Solutions.....	221

# About This Book

## What Does This Book Cover?

This book is designed to fulfill two purposes: one is to teach statistical concepts and the other is to show you how to perform statistical analysis using SAS Studio.

The book starts out with two short, introductory chapters describing statistical analysis (both descriptive and inferential) and experimental design. Following the introductory chapters are several chapters that show you how to download and install SAS University Edition (along with SAS Studio) and how to navigate your way around SAS Studio tasks. There is one chapter on descriptive statistics, summarizing data both in table and graphical form. The remainder of the book describes most of the statistical tests that you will need for an introductory course in statistics.

## Is This Book for You?

As the title suggests, this book is intended for someone with little or no knowledge of statistics and SAS, but it is also useful for someone with more statistical expertise who might not be familiar with SAS. One of the important points for beginners or people with more extensive knowledge of statistics, is a discussion of the assumptions that need to be satisfied for a particular statistical test to be valid. That is especially important because with SAS Studio tasks, anyone can click a mouse and perform very advanced statistical tests.

## What Should You Know about the Examples?

Because you can download all of the programs and data sets used in this book from the SAS website, you can run any or all of the programs yourself to better understand how perform them.

### Example Code and Data

You can access the example code and data for this book by linking to its author page at <https://support.sas.com/cody>.

### SAS University Edition



This book is compatible with SAS University Edition. Although all the examples in the book were run using SAS University Edition, you can use SAS On Demand for Academics or SAS Studio.

## Where Are the Exercise Solutions?

Solutions to all the odd-numbered exercises are included at the end of the book. For those individuals who are not students, are working on their own, or are faculty members, please contact SAS Press for solutions to all of the exercises.

## We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit [sas.com/books](https://sas.com/books) to do the following:

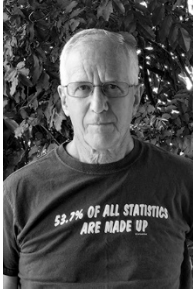
- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through [saspress@sas.com](mailto:saspress@sas.com) or [https://support.sas.com/author\\_feedback](https://support.sas.com/author_feedback).

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: [sas.com/books](https://sas.com/books).



## About The Author



Ron Cody, EdD, is a retired professor from the Rutgers Robert Wood Johnson Medical School who now works as a national instructor for SAS and as an author of books on SAS and statistics. A SAS user since 1977, Ron's extensive knowledge and innovative style have made him a popular presenter at local, regional, and national SAS conferences. He has authored or co-authored numerous books, as well as countless articles in medical and scientific journals.

Learn more about this author by visiting his author page at <http://support.sas.com/cody>. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.



# Chapter 9: Comparing More Than Two Means (ANOVA)

Introduction .....	93
Getting an Intuitive Feel for a One-Way ANOVA.....	94
Performing a One-Way Analysis of Variance.....	94
Performing More Diagnostic Plots.....	102
Performing a Nonparametric One-Way Test .....	104
Conclusion .....	109
Chapter 9 Exercises .....	109

## Introduction

When you want to compare means in a study where there are three or more groups, you cannot use multiple  $t$  tests. In the old days (even before my time!), if you had three groups (let's call them A, B, and C), you might perform  $t$  tests between each pair of means (A versus B, A versus C, and B versus C). With four groups, the situation gets more complicated; you would need six  $t$  tests (A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D). Even though no one does multiple  $t$  tests anymore, it is important to understand the underlying reason why this is not statistically sound.

Suppose you are comparing four groups and performing six  $t$  tests. Also, suppose that the null hypothesis is true, and all the means come from populations with equal means. If you perform each  $t$  test with  $\alpha$  set at .05, there is a probability of .95 that you will make the correct decision—that is, to fail to reject the null hypothesis in each of the six tests. However, what is the probability that you will reject at least one of the six null hypotheses? To spare you the math, the answer is about .26 (or 26% if that is easier to think about). This is called an "experimentwise" type I error. Remember, a type I error is when you reject the null hypothesis (claim the samples come from populations with different means—"the drug works") when you shouldn't. So, instead of your chance of reporting a false positive result being .05, it is really .26.

To prevent this problem, statisticians came up with a single test, called **analysis of variance** (abbreviated ANOVA). The null hypothesis is that all the means come from populations with equal means; the alternative hypothesis is that there is at least one mean that is different from the others. You either reject or fail to reject the null hypothesis, and there is one  $p$ -value associated with the test. If you reject the null hypothesis, you can then investigate pairwise differences using methods that control the experimentwise type I error.

## Getting an Intuitive Feel for a One-Way ANOVA

Before we get into the details of running and interpreting ANOVA tables, let's get an intuitive feel for how this analysis works. Suppose you have three groups of subjects (A, B, and C) and you collected the following data:

Group	A	B	C
	50	78	20
	45	80	15
	55	82	26
<b>Means</b>	<b>50</b>	<b>80</b>	<b>20</b>

You see the means in groups A, B, and C are 50, 80, and 20 respectively. They seem pretty far apart. But, what does "far apart" mean? In this case, they are far apart compared to the scores within each group (which seem very close to the group mean). This might lead you to think that there is a significant difference between the groups.

In English, when there are more than two groups, proper grammar is to say "among", not "between". However, the terms "within" and "between" have been used to describe variances in ANOVA designs since they were first developed, and most textbooks have kept with these terms.

You can skip this next paragraph if you want—it describes how ANOVA works in more detail.

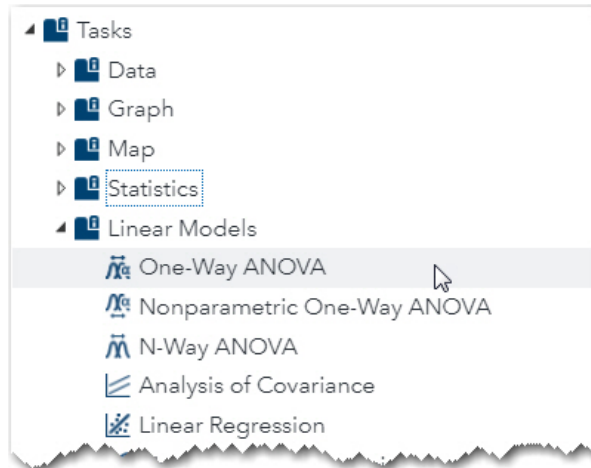
You can estimate the population variance by looking at the scores within a group or by using the group means to estimate the variance. If the null hypothesis were true (all the sample means come from populations with equal means), these two estimates of variance would be about the same and the ratio of the between-group variance to the within-group variance (called an **F value**) would be close to 1. If there were significant differences between the groups, the variance estimate computed by using the group means would be larger than the variance estimate computed by looking at the scores within a group. In this case, the F ratio would be greater than 1.

## Performing a One-Way Analysis of Variance

We can use the data set called Reading (in the STATS library), containing data on reading speeds of males and females, as well as three different methods that might improve reading speeds of the test subjects, to demonstrate a one-way ANOVA.

Start by choosing the task One-Way ANOVA from the statistics Tasks ► Linear Models task list. This brings up the following screen:

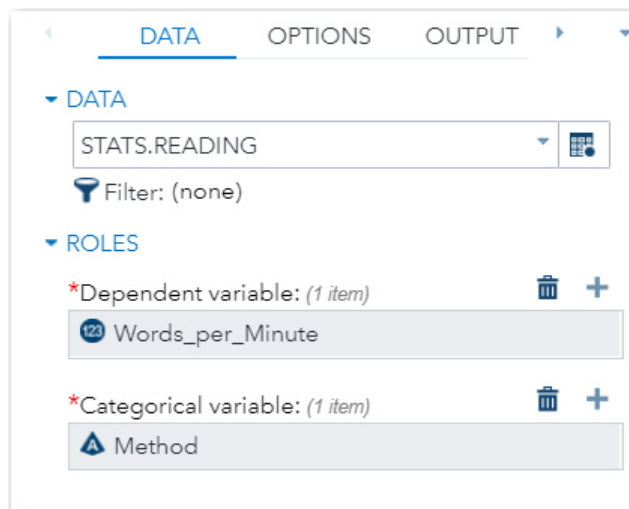
**Figure 9.1: Selecting One-Way ANOVA from Linear Models List**



To begin, double-click the One-Way ANOVA task tab.

Choose the data set Reading in the STATS library. This library was created when you ran the program Create\_Datasets.sas. Choose Words\_per\_Minute as the Dependent variable and Method as the Categorical (independent) variable, as shown in Figure 9.2.

**Figure 9.2: DATA Tab for One-Way ANOVA**



Once you have completed the DATA screen, click the OPTIONS tab to see the following.

**Figure 9.3: OPTIONS for One-Way ANOVA (Top Portion)**

The screenshot shows the SAS Studio interface for the OPTIONS tab of a One-Way ANOVA procedure. The tabs at the top are DATA, OPTIONS, OUTPUT, and INFORMATION. The OPTIONS tab is active. Under the 'HOMOGENEITY OF VARIANCE' section, the 'Test:' dropdown menu is set to 'Levene'. Below this, the checkbox for 'Welch's variance-weighted ANOVA' is unchecked. Under the 'COMPARISONS' section, the 'Comparisons method:' dropdown menu is set to 'Tukey'. At the bottom, the 'Significance level:' is set to '0.05'.

One of the assumptions for performing an analysis of variance is that the variances in each of the groups are equal. The **Levene test** is used to determine if this assumption is reasonable. If this test is significant (meaning the variances are not equal), you may choose to ignore it if the differences are not too large. (ANOVA is said to be robust to the assumption of equal variance, especially if the sample sizes are similar). If you want to account for unequal variances, click the box for Welch's variance-weighted ANOVA.

Multiple comparisons are methods that you use in order to determine which pairs of means are significantly different. There are several choices for these tests. The default is Tukey, a popular choice. Later in this chapter, you will see another multiple comparison test called SNK (Student-Newman-Keuls). You probably want to leave the significance level at .05.

It's time to run the procedure. Click the Run icon to produce the tables and graphs.

The first section of output displays class-level information. Don't ignore this! Make sure that the number of levels is what you expected (data errors can cause the program to believe there are more levels than there are). Also, pay attention to the number of observations read and used. This is important because any missing values on either the dependent variable (Words\_per\_Minute) or categorical variable (Method) will result in that observation being omitted from the analysis. A large proportion of missing values in the analysis can lead to bias—subjects with missing values might be different in some way from subjects without missing values (that is, missing values might not be random).

Figure 9.4: Class-Level Information

Class Level Information		
Class	Levels	Values
Method	3	A B C

Number of Observations Read	120
Number of Observations Used	120

There are three levels for Method (A, B, and C) and there are no missing values (because the number of observations read is the same as the number of observations used). It's time to look at your ANOVA table (Figure 9.5 below).

Figure 9.5: ANOVA Table

Dependent Variable: Words_per_Minute					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	16856.53834	8428.26917	20.84	<.0001
Error	117	47324.92840	404.48657		
Corrected Total	119	64181.46674			

You can look at the F test and *p*-values in the ANOVA table, but you must remember that you also need to look at several other parts of the output to determine if the assumptions for the test are satisfied. You will see in the diagnostic tests that follow that the ANOVA assumptions were satisfied, so let's go ahead and see what conclusions you can draw from the ANOVA table and the tables that follow.

Notice that the model has 2 **degrees of freedom** (because there were 3 levels of the independent variable, and the degrees of freedom is the number of groups minus 1). The mean squares for the model and error terms tell you the between-group variance and the within-group variance. The ratio of these two variances, the F value, is 20.84 with a corresponding *p*-value of less than .0001. A result such as this is often referred to as "highly significant." Remember, the term "significant" means that the probability of falsely rejecting the null hypothesis is smaller than a pre-determined value. It doesn't necessarily mean that the differences are significant in the common usage of the word, that is, important.

To graphically display the distribution of reading speed (Words\_per\_Minute) in the 3 groups, the one-way ANOVA task produces a box plot (Figure 9.6). The line in the center of the box represents the median, and the small diamond represents the mean. Notice that the means, as well as the medians, of the three groups are not very different. Why then were the results so

highly significant? The reason is the large sample size (120). Large sample sizes give you high power to see even small differences.

**Figure 9.6: Box Plot for Words\_per\_Minute by Method**

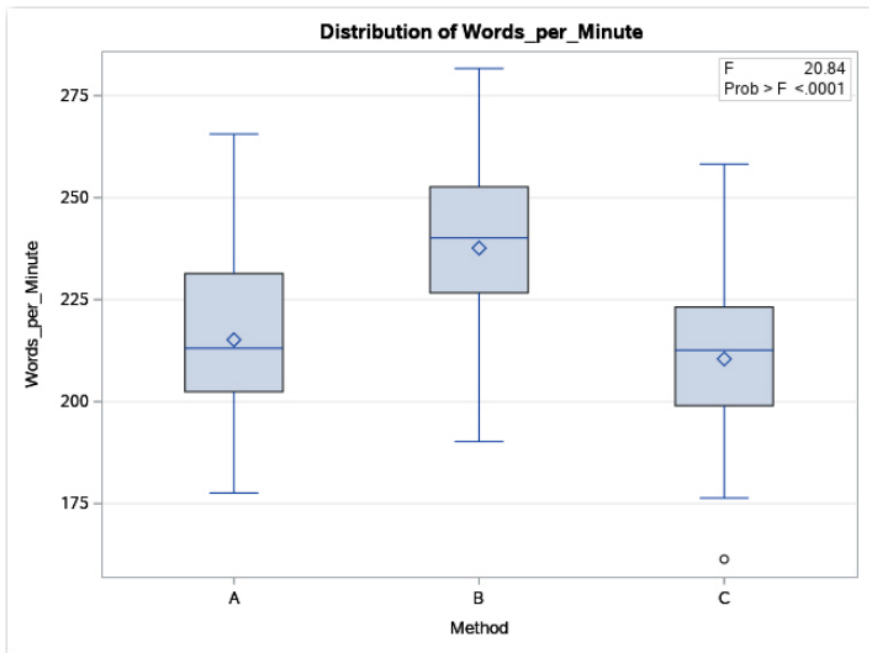


Figure 9.7 shows the results for Levin's test of homogeneity of variance. Here, the null hypothesis is that the variances are equal. Because the  $p$ -value is .9425, you do not reject the null hypothesis of equal variance.

**Figure 9.7: Levene's Test for Homogeneity of Variance**

Levene's Test for Homogeneity of Words_per_Minute Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Method	2	36892.2	18446.1	0.06	0.9425
Error	117	36445775	311502		



Figure 9.8 show the means and standard deviations for the three groups.

**Figure 9.8: Means and Standard Deviations for the Three Groups**

Level of Method	N	Words_per_Minute	
		Mean	Std Dev
A	40	215.161350	20.7279745
B	40	237.627783	19.7155850
C	40	210.470850	19.8772856

Because this is a one-way model, the least square means shown in Figure 9.9 (below) are equal to the means computed by adding up all the values within a group and dividing by the number of subjects in that group. In unbalanced models with more than one factor, this might not be the case.

Below the table showing the three means, you see  $p$ -values for all of the pairwise differences. Each of the three reading methods in the top table in the figure has what is labeled as the LSMEAN Number. In the table of  $p$ -values, the LSMEAN number is used to identify the groups. The intersection of any two groups displays the  $p$ -value for the difference. For example, group 1 (Method A) and group 2 (Method B) show a  $p$ -value of less than .0001. The  $p$ -value for the difference of Method A (1) and Method C (3) is .5514 (not significant).

**Figure 9.9: Least Square Means**

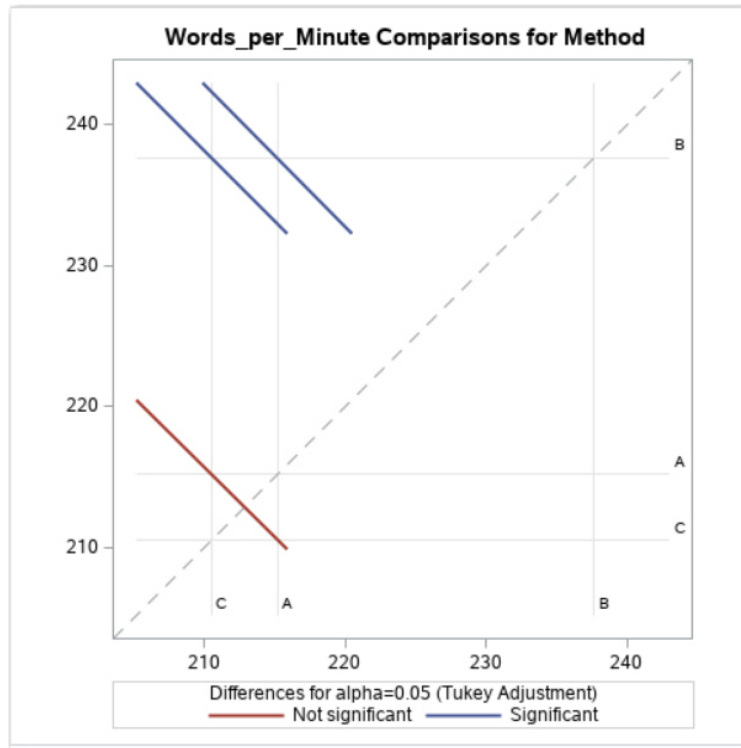
Least Squares Means			
Adjustment for Multiple Comparisons: Tukey			
Method	Words_per_Minute LSMEAN	LSMEAN Number	
A	215.161350	1	
B	237.627783	2	
C	210.470850	3	

Least Squares Means for effect Method			
Pr >  t  for H0: LSMean(i)=LSMean(j)			
Dependent Variable: Words_per_Minute			
i/j	1	2	3
1		<.0001	0.5514
2	<.0001		<.0001
3	0.5514	<.0001	

Figure 9.10 shows a very clever way to display pairwise differences. At the intersection of any two groups, you see a diagonal line representing a 95% confidence interval for the difference between the two group means. If the interval crosses the main diagonal line (that represents no difference), the two group means are not significantly different at the .05 level. To make this clearer, significant differences are seen in the two top diagonal lines representing C versus B and A versus B (they don't cross the dotted line) and the diagonal line at the bottom left of the diagram representing C versus A, indicates a non-significant difference. By the way, the name **diffogram** is used to describe this method of displaying pairwise differences.

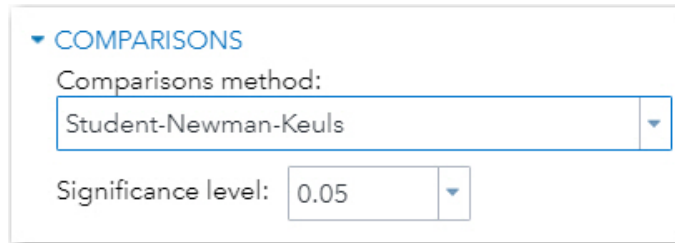
Figure 9.10: Pairwise Comparison of Means



All of the previous figures were generated by the choices that you made in the DATA and OPTIONS tabs. There is an alternative method of determining pairwise differences called the **Student-Newman-Keuls (SNK) test** (also referred to in some texts as just Newman-Keuls). The SNK test is similar to the Tukey test in that it shows group means and which pairs of means are different at the .05 level. The Tukey test has the advantage of computing  $p$ -values for each pair of means as well as a confidence interval for the differences. The SNK test can do neither of these two things but has a slightly higher power to detect differences.

To request the SNK multiple comparison test, select Student-Newman-Keuls from the list under the Comparisons tab as shown in Figure 9.11

**Figure 9.11: Requesting an SNK Multiple Comparison Test**



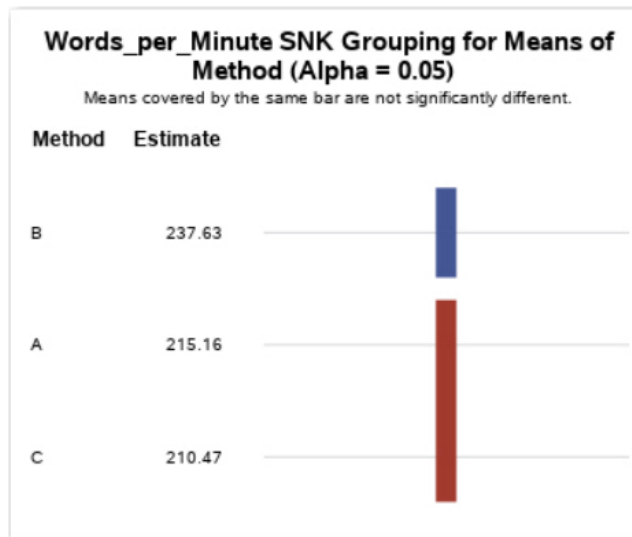
▼ **COMPARISONS**

Comparisons method:  
 Student-Newman-Keuls ▼

Significance level: 0.05 ▼

The SNK display (Figure 9.12) shows the three means in order from highest to lowest. Notice the bars on the right side of the output. Any two means that share the same bar are not significantly different at the .05 level. You can see here that the mean reading speed for group B is the highest, and it is significantly different from the mean of group A and from group C. Because groups A and C share a single bar, these two means (215.16 and 210.47) are not significantly different from each other.

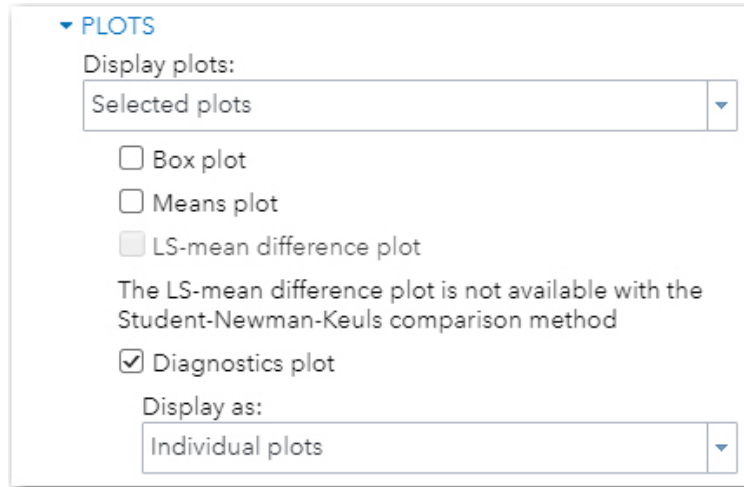
**Figure 9.12: Student-Newman-Keuls Pairwise Comparisons**



## Performing More Diagnostic Plots

Before we leave this section, let's look at a few diagnostic plots that you can choose on the PLOTS menu. In the pull-down list below Diagnostic Plots, you can select either Panel of Plots or Individual plots. The Panel option shows all the plots on a single page—the Individual option shows each diagnostic plot on a separate page. In this example, you decided to see individual plots.

**Figure 9.13: Requesting More Diagnostic Plots**



▼ PLOTS

Display plots:

Selected plots ▼

☐ Box plot

☐ Means plot

☐ LS-mean difference plot

The LS-mean difference plot is not available with the Student-Newman-Keuls comparison method

☒ Diagnostics plot

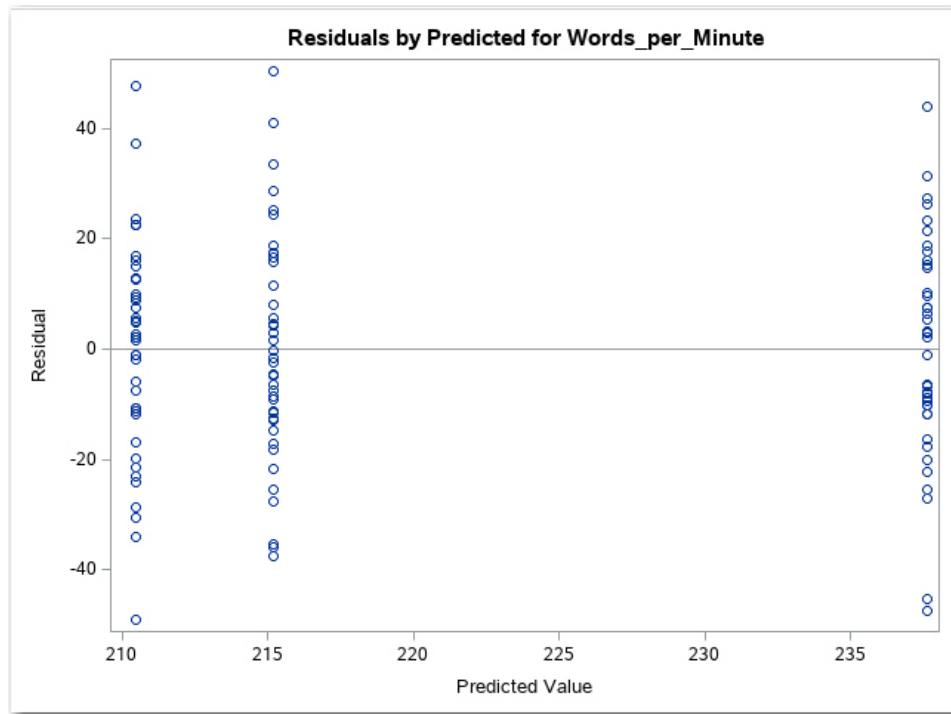
Display as:

Individual plots ▼

The next several plots are intended to help you decide if the ANOVA assumptions were satisfied and to graphically show you information about the three means and the distribution of scores in each of the three groups.

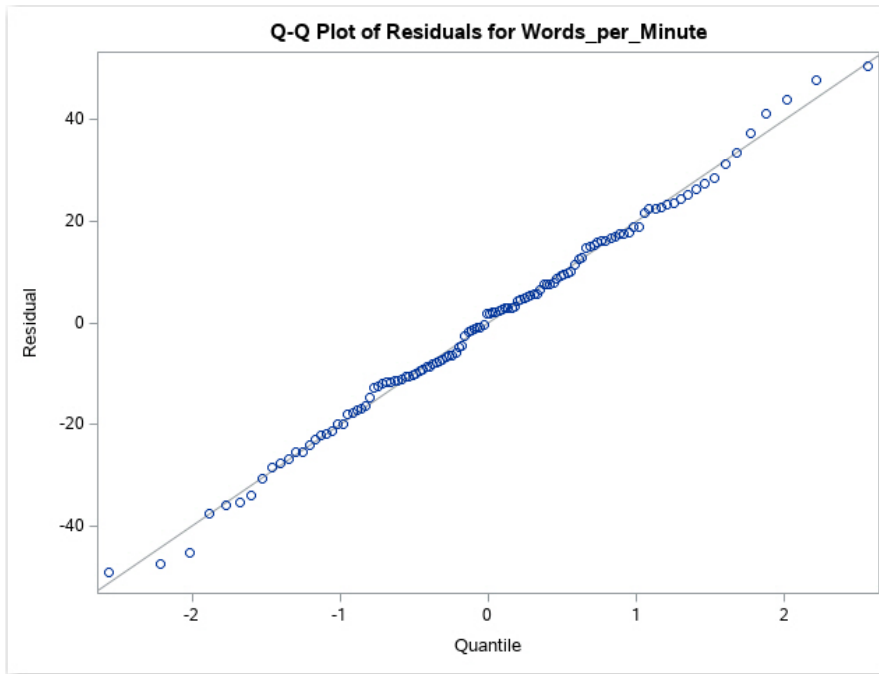
The figures shown below were selected from a larger set of plots produced by the one-way ANOVA task.

The plot shown in Figure 9.14 shows the residuals (the differences between the mean of each group and each individual score in that group). There are actually two residual plots produced by the one-way task. One (shown here) displays the residuals as actual scores (words-per-minute, in this example). Another residual plot (not shown) displays the residuals as  $t$  scores (the number of standard deviations above or below the mean of the group). Both plots look very similar. You also see the predicted values (means of each group) shown on the X axis.

**Figure 9.14: Residuals by Predicted Values**

Notice that the residuals are spread out equally above and below the zero value for each of the three groups. This is another way to see that the variances in the three groups are not significantly different (as shown by the Levene Test).

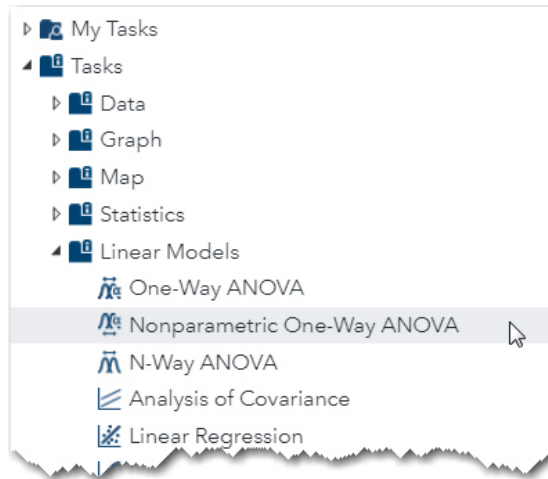
One of the assumptions for running a one-way ANOVA is that the errors (the residuals are estimates of these errors) are normally distributed. You have seen Q-Q plots earlier in this book, so you remember that data values that are normally distributed appear as a straight line on a Q-Q plot. The plot shown in Figure 9.15 shows small deviations from a straight line, but not enough to invalidate the analysis.

**Figure 9.15: Q-Q Plot of the Residuals**

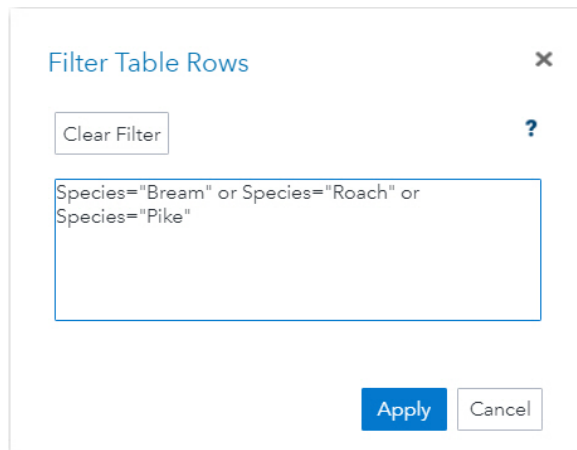
## Performing a Nonparametric One-Way Test

If you feel that the distribution assumptions are not satisfied by your data, another statistical task, Nonparametric One-Way ANOVA, provides a host of alternate tests. To demonstrate this, let's go back to the SASHELP data set called Fish and compare the weights of three species of fish.

Start out by selecting Nonparametric One-Way ANOVA found under the Linear Models in the Tasks list (Figure 9.16).

**Figure 9.16: Select Nonparametric One-Way ANOVA from the Linear Models Tab**

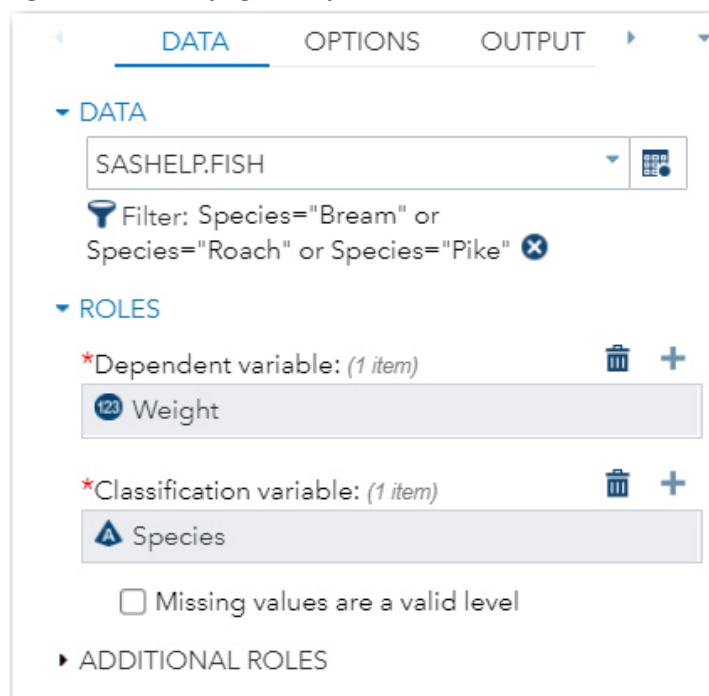
Next, you want to create a filter to select three species: Bream, Roach, and Pike. The expression that you need to write is shown in Figure 9.17 below.

**Figure 9.17: Creating a Filter for Three Species**

The names of the three species need to be placed in either single or double quotation marks because Species is a character variable. Once you apply this filter, only the three species will be used in the analysis.

On the DATA Tab, select Weight as the Dependent variable and Species as the Classification variable (Figure 9.18).

**Figure 9.18: Identifying the Dependent and Classification Variables**



Next, click the OPTIONS tab. For this example, you are using all the default values except for a request for Pairwise multiple comparison analysis (asymptotic only).



Figure 9.19: Options for the Analysis

DATA   **OPTIONS**   OUTPUT

▶ PLOTS

▼ TESTS

Tests:

Asymptotic tests ▼

▼ Location Differences

☒ Wilcoxon scores

☐ Median scores

☐ Van der Waerden scores

☐ Savage scores

▶ Scale Differences

▶ Location and Scale Differences

▼ Additional Tests

☐ Empirical distribution function tests, including Kolmogorov-Smirnov and Cramer-von Mises tests, or the Kuiper test (for two-sample data)

☒ Pairwise multiple comparison analysis (asymptotic only)

Now that you have entered your choices on the DATA and OPTIONS tabs, it's time to run the task. The first part of the output is shown in Figure 9.20.

Figure 9.20: Results from the Wilcoxon Rank Sum Test

Wilcoxon Scores (Rank Sums) for Variable Weight Classified by Variable Species					
Species	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Bream	34	1580.00	1224.0	86.852273	46.470588
Roach	20	224.50	720.0	78.206158	11.225000
Pike	17	751.50	612.0	74.192876	44.205882
Average scores were used for ties.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
40.2791	2	<.0001

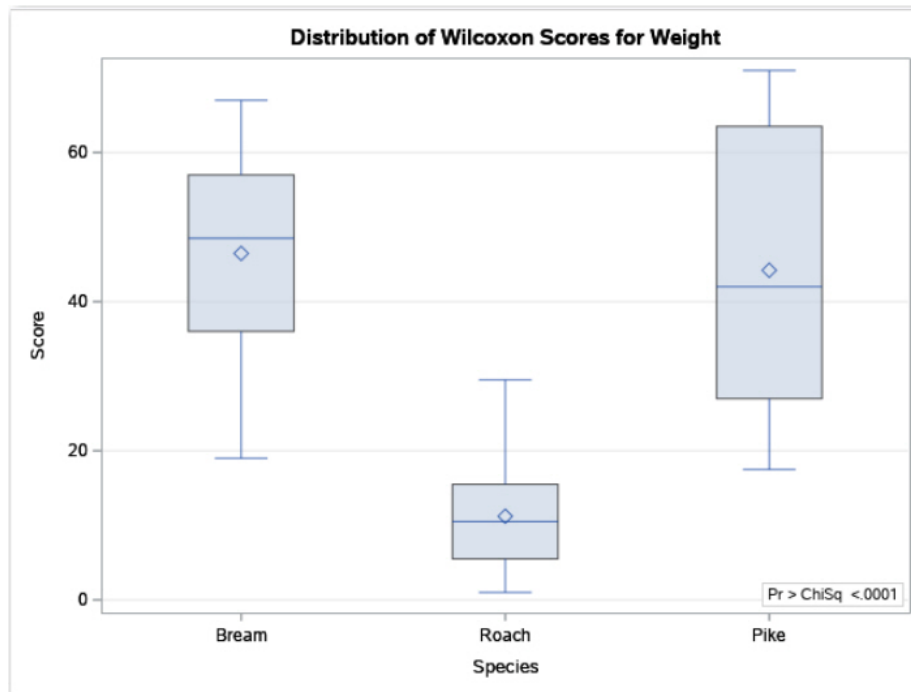
To perform the Wilcoxon Rank Sum Test, all the scores are sorted in order from the lowest score to the highest score. You then assign a rank to each score: the lowest score is rank 1, the next

highest score is rank 2, and so on. If several scores are equal, the Wilcoxon test assigns the mean rank to each of the values (don't worry about this detail for now). Along with the fish weights, you also know the species associated with each rank. There is a total of 71 fish weights ( $34 + 20 + 17$ ), so the ranks range from 1 to 71. If the null hypothesis is that the fish are all about the same weight, you would expect the sum of ranks for each species to be about the same. You use this idea to form your null hypothesis. If one or more of the fish species has a very high or low sum of ranks, you might expect that there are differences in weight based on species.

The table above shows the sum of ranks for each of the fish species and the expected value if the null hypothesis is true. Notice that the sum of Scores for Roach (224.5) is much lower than the sums for Bream or Pike, making you suspect that Roach are typically lighter than either Bream or Pike. To decide whether you should reject the null hypothesis that all the three species have equal weights, you look at the  $p$ -value at the bottom of Figure 9.20. Because the  $p$ -value is shown as  $<.0001$ , you reject the null hypothesis and conclude that one or more pairs of means are significantly different. But, which pairs of fish species are different? We will answer that question in a minute.

The next part of the output shows box plots for each fish species. Box plots are appropriate for this display because you are conducting a nonparametric test.

**Figure 9.21: Box plots for Each Species**



This plot shows that Roach seem to be lighter than either Bream or Pike. Just as you did a multiple comparison test in the ANOVA test (Tukey or SNK), there is an equivalent multiple

comparison nonparametric test. Figure 9.22 shows significant differences between Bream versus Roach and Roach versus Pike. Bream and Pike are not statistically different ( $p = .8900$ ).

Figure 9.22: Pairwise Comparisons

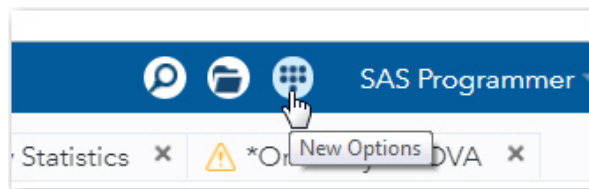
Pairwise Two-Sided Multiple Comparison Analysis			
Dwass, Steel, Critchlow-Fligner Method			
Variable: Weight			
Species	Wilcoxon Z	DSCF Value	Pr > DSCF
Bream vs. Roach	5.9671	8.4388	<.0001
Bream vs. Pike	0.4599	0.6504	0.8900
Roach vs. Pike	-4.9544	7.0066	<.0001

## Conclusion

You have seen how to conduct a one-way analysis of variance as well as a Wilcoxon nonparametric test. You have also seen ways to determine if the two assumptions for a one-way ANOVA (normally distributed data and homogeneity of variance) are met.

## Chapter 9 Exercises

1. List the first 10 observations from the High\_School data set found in the STATS library (this was created when you ran the Create\_Dataset.sas program in the download package). Conduct a one-way ANOVA comparing the variable Vocab\_Score (a measure of vocabulary skill) by Grade (Freshman, Sophomore, Junior, and Senior). Be sure to run a Tukey multiple comparison test to determine which grades are different from each other (or none).
2. Repeat exercise 1, except this time, use the variable English\_Grade as the dependent variable.
3. The data set Salary\_Formatted in the STATS library contains variables Gender, Age\_Group and Weekly\_Salary. First, run the short program below to create a new variable (Gender\_Age) that creates four combinations of the two variables Gender and Age\_Group. (Hint: click on the new options icon and request a New Program.)



```

data Temp;
  set Stats.Salary_Formatted;
  length Gender_Age $ 6;
  Gender_Age = Cats(Gender, Age_Group);
run;

```

The first 10 observations in the Temp data set should look like this:

Obs	Gender	Age_Group	Education	Weekly_Salary	Salary	Gender_Age
1	Female	20-24	<HS	443	Below the Median	F20-24
2	Female	20-24	BA+	1574	Above the Median	F20-24
3	Female	45-54	<HS	1089	Below the Median	F45-54
4	Female	45-54	BA+	1861	Above the Median	F45-54
5	Male	20-24	<HS	939	Below the Median	M20-24
6	Male	20-24	<HS	872	Below the Median	M20-24
7	Male	20-24	<HS	819	Below the Median	M20-24
8	Male	20-24	<HS	873	Below the Median	M20-24
9	Male	20-24	<HS	970	Below the Median	M20-24
10	Male	20-24	<HS	680	Below the Median	M20-24

Using the Temp data set, run a one-way analysis using Weekly\_Salary as the dependent variable and Gender\_Age as the categorical variable.

- Using the data set Fish in the SASHELP library, compare the Width (not the Weight) of three species of fish; Perch, Roach, and Pike. Do this using both parametric and nonparametric methods. You will need to create a filter that reads:

```
Species='Perch' or Species='Roach' or Species='Pike'
```

- Using the data set Heart in the SASHELP library, run a nonparametric ANOVA using Cholesterol as your dependent variable and DeathCause as your classification variable. Include an option for multiple comparisons. Which, if any, of the causes of death had significant differences in cholesterol?
- Using the data set Cars in the SASHELP library, compare the Horsepower for each Type of car. Use a filter (Type ne 'Hybrid') to eliminate hybrids because there are so few of them.
- Using the data set Cars in the SASHELP library, compare the Weight of four-, six-, and eight-cylinder cars. Use a filter to restrict the variable Cylinders to values of 4, 6, or 8. Hint: the filter expression should read: Cylinders = 4 or Cylinders=6 or Cylinders=8. An interesting alternative is: Cylinders IN (4,6,8).