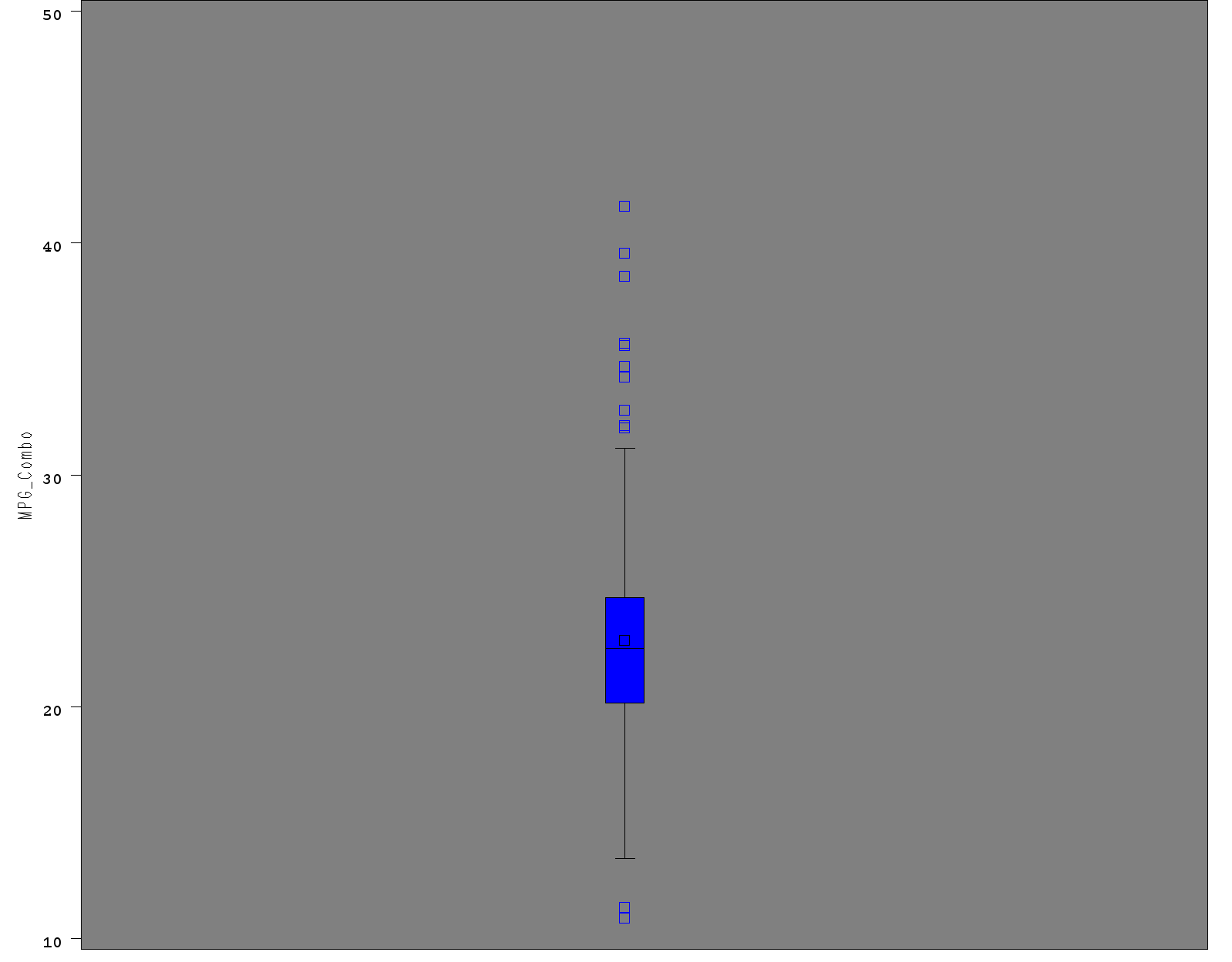
**Exercise 1: Descriptive Statistics**

1. Create a combined mpg variable called **MPG\_Combo** which combines 55% of the **MPG\_City** and 45% of the **MPG\_Highway**. Obtain a box plot for **MPG\_Combo** and comment on what the plot tells us about fuel efficiencies.

The below box plot shows MPG\_Combo mean (22.88) to be slightly greater than median (22.5). This shows that the distribution is right-skewed. There is presence of many outliers above 30 that inflate the average MPG

Since the median is less than mean and represents 50% of the data, hence there are more observations with low fuel efficiency values.

***Distribution analysis of: MPG\_Combo***

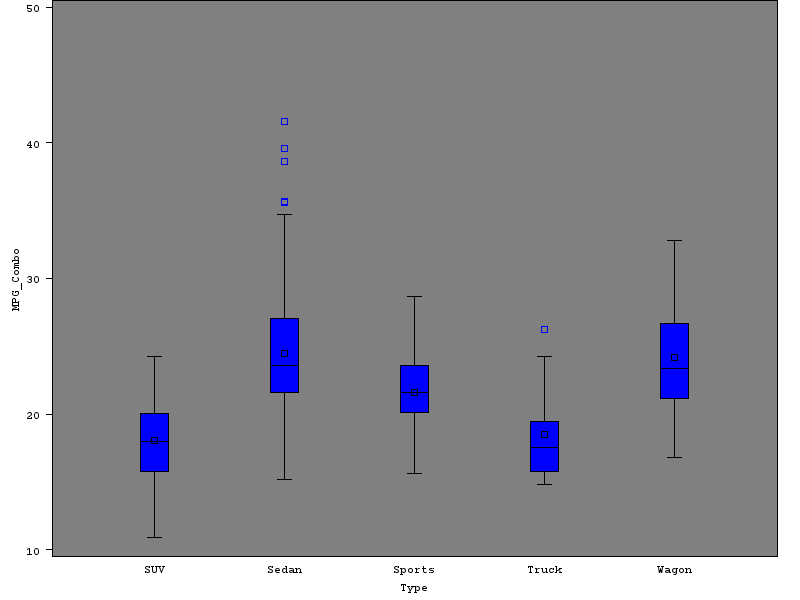


1. **Obtain box plots for MPG\_Combo by Type and comment on any differences you notice between the different vehicle types combined fuel efficiency.**

Sedan and Wagon appear to be more Fuel efficient(Sedan being the most) than the rest of the vehicle types since their mean is around 24, which is greater than that of others.

Of the two SUV and truck that have less fuel efficiency, SUV has the least fuel efficiency of all.

Sports type has moderate fuel efficiency



1. **Obtain basic descriptive statistics for Invoice for all vehicles. Comment on any general features and statistics of the data. Use visual and quantitative methods to comment on whether an assumption of Normality would be reasonable for Invoice variable.**

|  |  |
| --- | --- |
| ***The UNIVARIATE Procedure*** | |
| ***Variable: Invoice*** |

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 425 | **Sum Weights** | 425 |
| **Mean** | 30096.48 | **Sum Observations** | 12791004 |
| **Std Deviation** | 17677.3562 | **Variance** | 312488924 |
| **Skewness** | 2.82591763 | **Kurtosis** | 13.8776543 |
| **Uncorrected SS** | 5.17459E11 | **Corrected SS** | 1.32495E11 |
| **Coeff Variation** | 58.735627 | **Std Error Mean** | 857.477729 |

**Descriptive Statistics**

* **Number of Observations** for Invoice = 425
* **Measure of Central tendency**: Mean = 30096.48, Median = 25672. Since Mean > median, data is skewed to the right(positively skewed)
* **Spread of Data:** Range, Variance and Std Deviation as shown below
* **Shape of distribution:**

1. **Skewness** = 2.82, which shows asymmetry in the distribution. Since it is a positive value(2.82 > 0) the data is positively skewed or right skewed
2. **Kurtosis** = 13.87. Since the value is greater than 3 the distribution has longer and fatter tails than normal distribution meaning data is concentrated at the tails showing the presence of outliers. Moreover, the peak is higher and sharper when compared to normal distribution

| **Basic Statistical Measures** | | | |
| --- | --- | --- | --- |
| **Location** | | **Variability** | |
| **Mean** | 30096.48 | **Std Deviation** | 17677 |
| **Median** | 25672.00 | **Variance** | 312488924 |
| **Mode** | 14207.00 | **Range** | 163685 |
|  |  | **Interquartile Range** | 16804 |

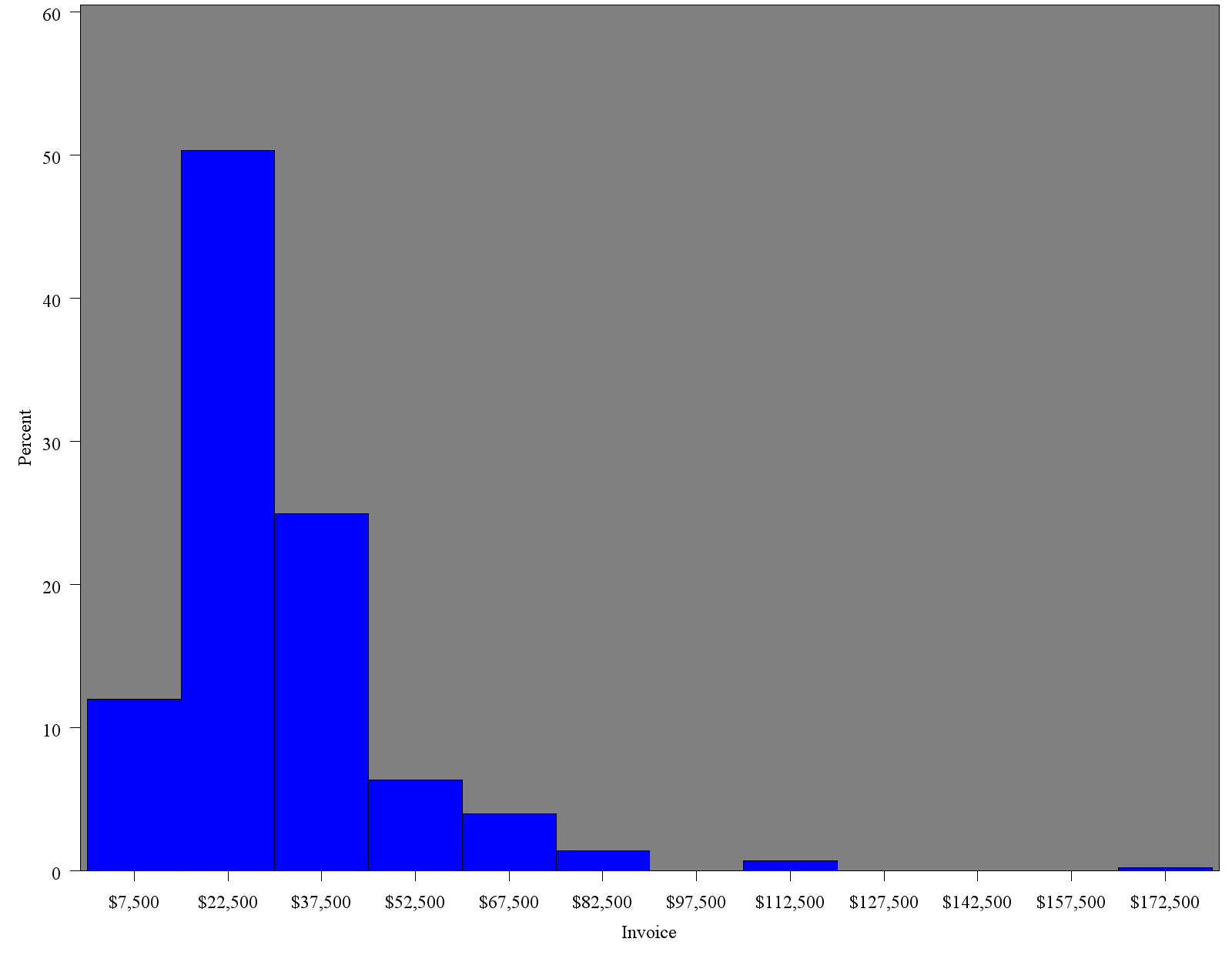
**Quantitative Method for testing Normality: Shapiro-Wilk Test**

Shapiro-Wilk Test resulted in a p value <0.0001. Since the probability is significant at p < 0.05, the Normality assumption is not met.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.77353 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.140604 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 3.393462 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 20.06351 | **Pr > A-Sq** | <0.0050 |

**Visual/Qualitative Method for Assumption of normality: Histogram**

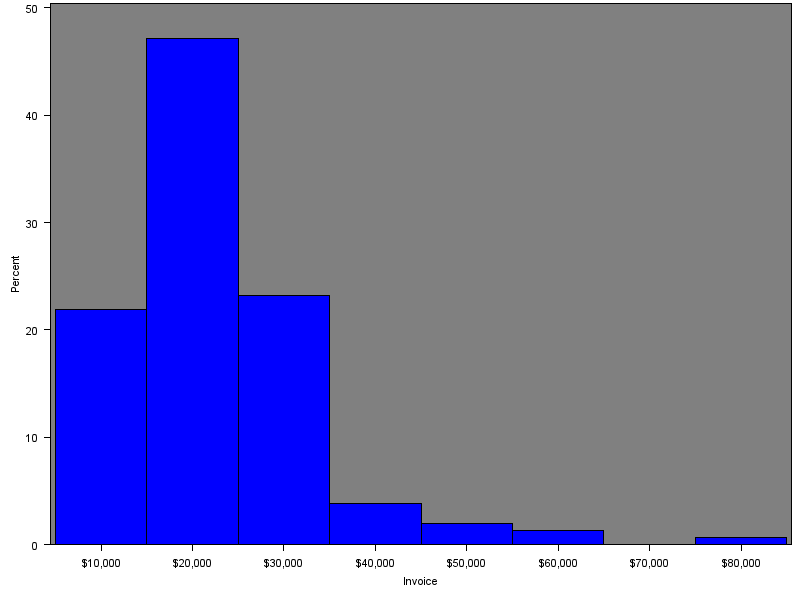
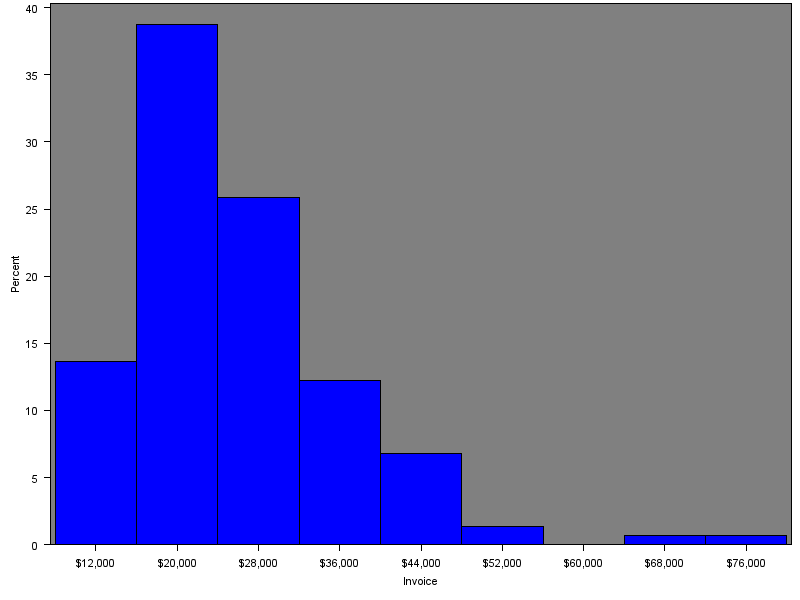
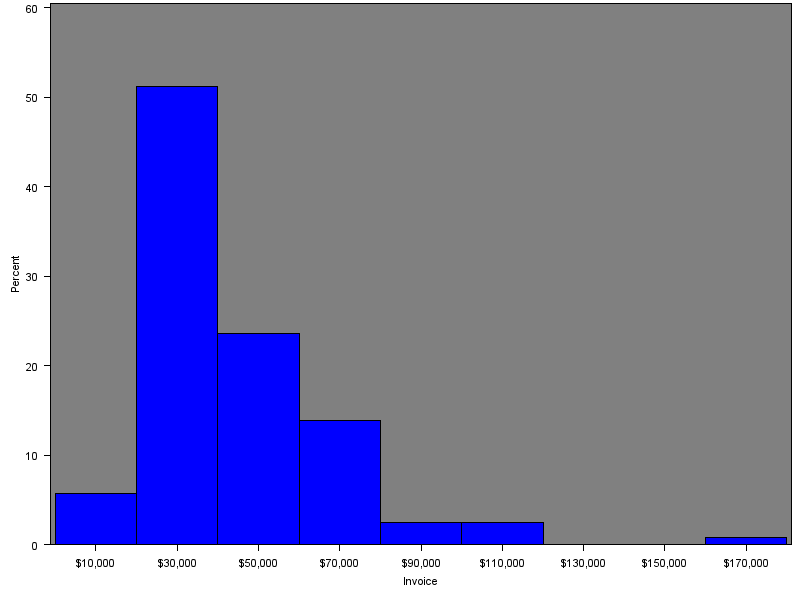
The histogram below shows the distribution is right-skewed which is also seen through descriptive statistics table mean (which is 30096) > median (which is 25672). Hence assumption of normality is not met. Also as predicted by the p value the data does not follow normal distribution.



1. **Use visual and quantitative methods to comment on whether an assumption of normality would be reasonable for Invoice variable by Origin. (i.e., check normality of Invoice from i) Europe, ii) Asian, and iii) USA cars**

**Normality of invoice from i) Europe, ii) Asian, and iii) USA respectively:**

**Visual method:** Histograms for all three invoice distributions show that it is a positively/right-skewed so normality assumption is not met**.**



**Quantitative method:** The P value from Shapiro-Wilk test is significant at p < 0.05 so the invoice data of Europe, Asia or USA is not normally distributed. Hence normality assumption is not met.

**Europe**

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.798086 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.175505 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 1.116219 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 6.068997 | **Pr > A-Sq** | <0.0050 |

**Asia**

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.846964 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.115286 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.601988 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 4.140188 | **Pr > A-Sq** | <0.0050 |

**USA**

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.892216 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.111942 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.552353 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 3.252361 | **Pr > A-Sq** | <0.0050 |

**Exercise 2: Hypothesis Testing**

1. **Perform a hypothesis test of whether cars originated in Europe have different invoice price than Asian cars, and state your conclusions. If they have significantly different invoice, which group has greater mean or median? Recall the test of normality from the above Problem when choosing your test.**

Shapiro-Wilks test proves that both Europe and Asia have significantly different invoice prices hence we reject null hypothesis in both cases

Since for Europe and Asian cars invoice data distribution, the normality assumption is not met, we use non-parametric Wilcoxon Rank sum test. The P value is significant at p < 0.05 hence Europe and Asian cars have significantly different invoice, we reject the null hypothesis

We perform the median non-parametric one-way ANOVA. Europe cars invoice price has greater median than that of Asian cars.

1. **Repeat a) for Asian and USA cars**

Shapiro-Wilks test proves that both USA and Asia have significantly different invoice prices hence we reject null hypothesis

Since for USA and Asian cars invoice data distribution, the normality assumption is not met, we use non-parametric Wilcoxon Rank sum test. The P value is significant at p < 0.05 hence USA and Asian cars have significantly different invoice, we reject the null hypothesis

We perform the median non-parametric one-way ANOVA. Both medians are not significantly different, we retain null hypothesis

**Exercise 3: Analysis of Variance**

1. **Perform a one-way ANOVA for Cholesterol level with BP\_Status as the categorical predictor, test any assumptions of the model that should be tested (aside from normality and independence, which you do not need to test), comment on the significance of the model, and the variation described by the model.**

**Assumptions**:

* Homogeneity of variance can be assumed by Levene’s Test (F statistic is not significant at p < 0.05 hence we retain null hypothesis).
* Dependent variable(cholesterol) is continuous

The F value is significant at p < 0.05 so we reject null hypothesis (that there is no significant difference between group means) and conclude that there is at least one group mean that is significantly different than others. So the **model is significant**.

**Amount of Variation described by the model** is determined by R2:

| **R-Square** | **Coeff Var** | **Root MSE** | **Cholesterol Mean** |
| --- | --- | --- | --- |
| 0.024198 | 18.65388 | 43.47010 | 233.0351 |

Since is R2 is 2.4%, it means only 2.4% of the variation in response(cholesterol) is explained by the model (BP\_Status). Therefore, we can conclude that the data points fall relatively far from the model and the model doesn’t explain much of the variation in the data

1. **Comment on any significantly different cholesterol means as determined by the post-hoc test comparing all pairwise differences. Specifically explain what that tells us about differences in cholesterol levels across blood pressure status groups, like which group has the highest or lowest mean of Cholesterol**

From the Post-Hoc test (Tukey’s Studentized range test), it can be observed that:

Group pairs High-Normal, High-optimal, Normal-high, optimal-high have significantly different cholesterol means as the 95% Confidence interval does not include 0.

High BP\_status group has highest Cholesterol mean and optimal BP\_Status has lowest cholesterol mean.

**Exercise 4: Analysis of Variance**

1. **Perform a one-way ANOVA for mcv as a function of drinking group. Comment on statistical significance of**

**the model, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.**

**Predictor /independent variable**: drinkgroup

**Dependent variable**: mcv

**Assumptions**:

* Dependent variable (mcv) is continuous
* Normality and independence assumptions are valid
* Homogeneity of variance by Levene’s Test below shows F statistic is not significant at p < 0.05. This retains the Ho that group means have equal variances

| **Levene's Test for Homogeneity of mcv Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **drinkgroup** | 4 | 1940.3 | 485.1 | 0.33 | 0.8587 |
| **Error** | 340 | 501897 | 1476.2 |  |  |

All Assumptions being satisfied; the One-way ANOVA is conducted:

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 4 | 733.176652 | 183.294163 | 10.26 | <.0001 |
| **Error** | 340 | 6073.055232 | 17.861927 |  |  |
| **Corrected Total** | 344 | 6806.231884 |  |  |  |

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **drinkgroup** | 4 | 733.1766522 | 183.2941630 | 10.26 | <.0001 |

Ho: The mean differences between the 5 drinking groups are not statistically significant

Ha: At least one of the group means is significantly different from the others

**Statistical Significance of the Model**: The F statistic is significant at p < 0.05 which gives us enough evidence to reject Ho

We can conclude that at least one drinking group has significantly different cholesterol mean from others.

**Amount of Variation described by the model** is determined by R2:

| **R-Square** | **Coeff Var** | **Root MSE** | **mcv Mean** |
| --- | --- | --- | --- |
| 0.107721 | 4.687627 | 4.226337 | 90.15942 |

Since is R2 is 10.7%, it means only 10.7% of the variation in response(mcv) is explained by the model (drinking group). Therefore, we can conclude that the data points fall relatively far from the model and the model doesn’t explain much of the variation in the data

**Equal Variance Assumption**:

* Homogeneity of variance by Levene’s Test below shows F statistic is not significant at p < 0.05. This retains the Ho that group means have equal variances. Hence equal variance assumption can be trusted.

| **Levene's Test for Homogeneity of mcv Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **drinkgroup** | 4 | 1940.3 | 485.1 | 0.33 | 0.8587 |
| **Error** | 340 | 501897 | 1476.2 |  |  |

1. **Perform a one-way ANOVA for alkphos as a function of drinking group. Comment on statistical significance of the model, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.**

**Predictor /independent variable**: drinkgroup

**Dependent variable**: alkphos

**Assumptions**:

* Dependent variable (mcv) is continuous
* Normality and independence assumptions are valid
* Homogeneity of variance by Levene’s Test below shows F statistic is not significant at p < 0.05. This retains the Ho that group means have equal variances

| **Levene's Test for Homogeneity of alkphos Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **drinkgroup** | 4 | 1065624 | 266406 | 0.96 | 0.4293 |
| **Error** | 340 | 94305238 | 277368 |  |  |

All Assumptions being satisfied; the One-way ANOVA is conducted:

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 4 | 4945.6283 | 1236.4071 | 3.79 | 0.0050 |
| **Error** | 340 | 110857.5021 | 326.0515 |  |  |
| **Corrected Total** | 344 | 115803.1304 |  |  |  |

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **drinkgroup** | 4 | 4945.628301 | 1236.407075 | 3.79 | 0.0050 |

Ho: The mean differences between the 5 drinking groups are not statistically significant

Ha: At least one of the group means is significantly different from the others

**Statistical Significance of the Model**: The F statistic is significant at p < 0.05 which gives us enough evidence to reject Ho

We can conclude that at least one drinking group has significantly different cholesterol mean from others.

**Amount of Variation described by the model** is determined by R2:

| **R-Square** | **Coeff Var** | **Root MSE** | **alkphos Mean** |
| --- | --- | --- | --- |
| 0.042707 | 25.84372 | 18.05690 | 69.86957 |

Since is R2 is 4.2%, it means only 4.2% of the variation in response(alkphos) is explained by the model. Therefore, we can conclude that the data points fall relatively far from the model and the model doesn’t explain much of the variation in the data

**Equal Variance Assumption**:

* Homogeneity of variance by Levene’s Test below shows F statistic is not significant at p < 0.05. This retains the Ho that group means have equal variances. Hence equal variance assumption can be trusted.

| **Levene's Test for Homogeneity of alkphos Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Drinkgroup** | 4 | 1065624 | 266406 | 0.96 | 0.4293 |
| **Error** | 340 | 94305238 | 277368 |  |  |

1. **Perform post-hoc tests for models in a) and b). Comment on any similarities or differences you observe from their results.**

We have evidence to reject Ho in both a) and b) and through one-way ANOVA test, we only know that Ha (at least one mean is significantly different from the other) is true. But we do not know which drinking groups have significantly different cholesterol means.

Since assumption of homogeneity of variance is met, Post-Hoc test called Tukey’s Studentized range (HSD) test will specify which groups have significantly different means.

**Post-hoc test for model in a)**

Tukey's Studentized Range (HSD) Test for mcv

|  |  |
| --- | --- |
| **Note:** | This test controls the Type I experimentwise error rate. |

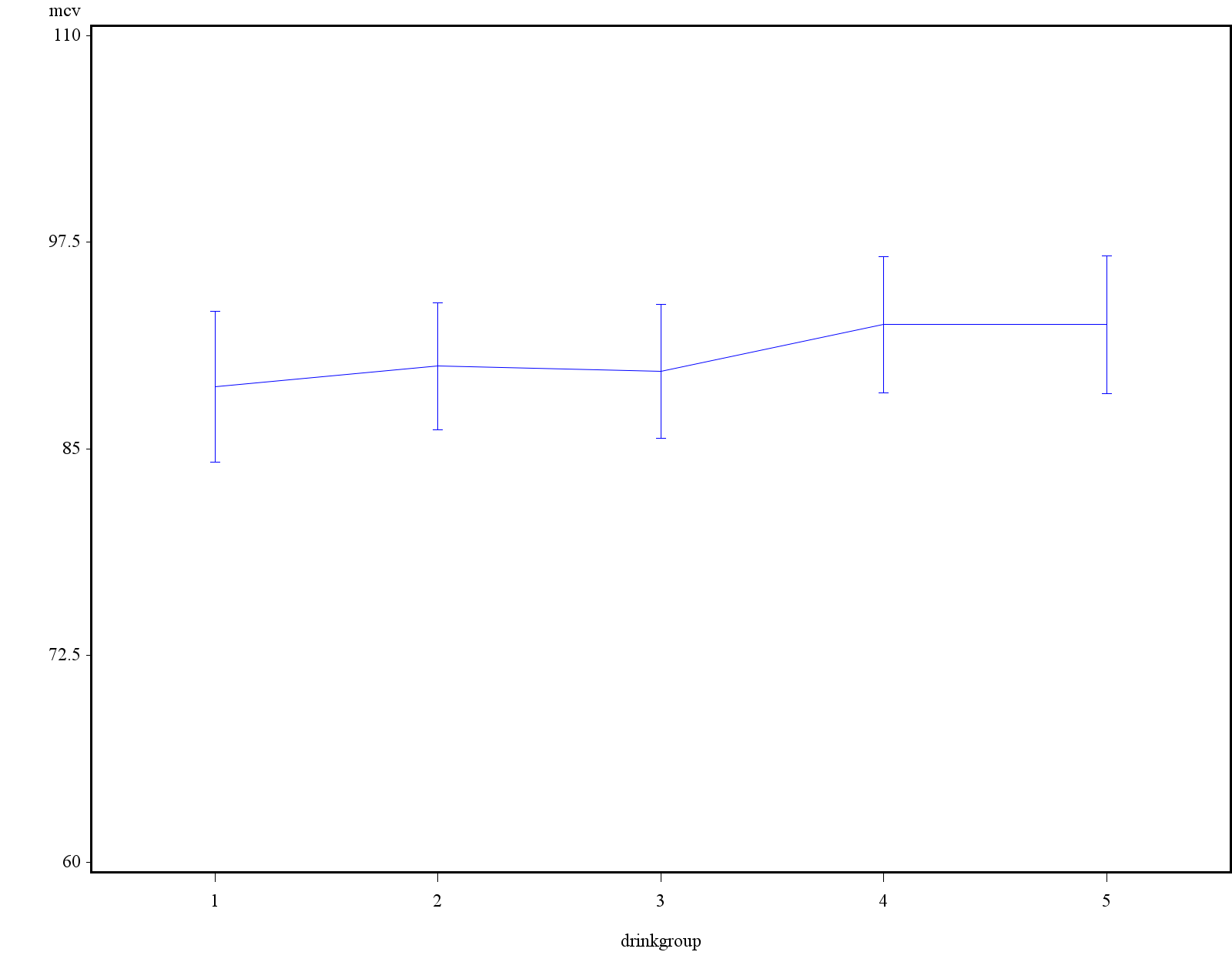
|  |  |
| --- | --- |
| **Alpha** | 0.05 |
| **Error Degrees of Freedom** | 340 |
| **Error Mean Square** | 17.86193 |
| **Critical Value of Studentized Range** | 3.87844 |

| **Comparisons significant at the 0.05 level are indicated by \*\*\*.** | | | | |
| --- | --- | --- | --- | --- |
| **drinkgroup Comparison** | **Difference Between Means** | **Simultaneous 95% Confidence Limits** | |  |
| **5 - 4** | 0.0014 | -2.8973 | 2.9001 |  |
| **5 - 2** | 2.5046 | -0.4922 | 5.5014 |  |
| **5 - 3** | 2.8079 | -0.0070 | 5.6228 |  |
| **5 - 1** | 3.7460 | 0.9991 | 6.4929 | \*\*\* |
| **4 - 5** | -0.0014 | -2.9001 | 2.8973 |  |
| **4 - 2** | 2.5032 | 0.3611 | 4.6453 | \*\*\* |
| **4 - 3** | 2.8065 | 0.9272 | 4.6858 | \*\*\* |
| **4 - 1** | 3.7446 | 1.9688 | 5.5204 | \*\*\* |
| **2 - 5** | -2.5046 | -5.5014 | 0.4922 |  |
| **2 - 4** | -2.5032 | -4.6453 | -0.3611 | \*\*\* |
| **2 - 3** | 0.3033 | -1.7240 | 2.3307 |  |
| **2 - 1** | 1.2415 | -0.6903 | 3.1732 |  |
| **3 - 5** | -2.8079 | -5.6228 | 0.0070 |  |
| **3 - 4** | -2.8065 | -4.6858 | -0.9272 | \*\*\* |
| **3 - 2** | -0.3033 | -2.3307 | 1.7240 |  |
| **3 - 1** | 0.9381 | -0.6974 | 2.5736 |  |
| **1 - 5** | -3.7460 | -6.4929 | -0.9991 | \*\*\* |
| **1 - 4** | -3.7446 | -5.5204 | -1.9688 | \*\*\* |
| **1 - 2** | -1.2415 | -3.1732 | 0.6903 |  |
| **1 - 3** | -0.9381 | -2.5736 | 0.6974 |  |

The above table compares multiple pairwise group means.

Group pairs 5-1, 4-2, 4-3, 4-1, 2-4, 3-4, 1-5, 1-4 have significant mean differences as the 95% Confidence interval does not include 0. This can also be viewed in the below means plot.

Drinking Group 5 and 4 have similar mcv mean values; **group 5 has the highest and group 1 has the least mean value**



**Post-hoc test for for model in b)**

| **Comparisons significant at the 0.05 level are indicated by \*\*\*.** | | | | |
| --- | --- | --- | --- | --- |
| **drinkgroup Comparison** | **Difference Between Means** | **Simultaneous 95% Confidence Limits** | |  |
| **5 - 1** | 12.573 | 0.837 | 24.309 | \*\*\* |
| **5 - 4** | 13.721 | 1.337 | 26.106 | \*\*\* |
| **5 - 2** | 15.218 | 2.414 | 28.022 | \*\*\* |
| **5 - 3** | 16.629 | 4.602 | 28.656 | \*\*\* |
| **1 - 5** | -12.573 | -24.309 | -0.837 | \*\*\* |
| **1 - 4** | 1.149 | -6.438 | 8.736 |  |
| **1 - 2** | 2.645 | -5.608 | 10.899 |  |
| **1 - 3** | 4.056 | -2.931 | 11.044 |  |
| **4 - 5** | -13.721 | -26.106 | -1.337 | \*\*\* |
| **4 - 1** | -1.149 | -8.736 | 6.438 |  |
| **4 - 2** | 1.497 | -7.656 | 10.649 |  |
| **4 - 3** | 2.907 | -5.122 | 10.937 |  |
| **2 - 5** | -15.218 | -28.022 | -2.414 | \*\*\* |
| **2 - 1** | -2.645 | -10.899 | 5.608 |  |
| **2 - 4** | -1.497 | -10.649 | 7.656 |  |
| **2 - 3** | 1.411 | -7.251 | 10.073 |  |
| **3 - 5** | -16.629 | -28.656 | -4.602 | \*\*\* |
| **3 - 1** | -4.056 | -11.044 | 2.931 |  |
| **3 - 4** | -2.907 | -10.937 | 5.122 |  |
| **3 - 2** | -1.411 | -10.073 | 7.251 |  |

Group pairs 5-1, 5-4, 5-2, 5-3, 1-5, 4-5, 2-5, 3-5 have significant means differences as the 95% Confidence interval does not include 0.

**Drinking group 5 has the highest and group 1 has the least alkphos mean value**