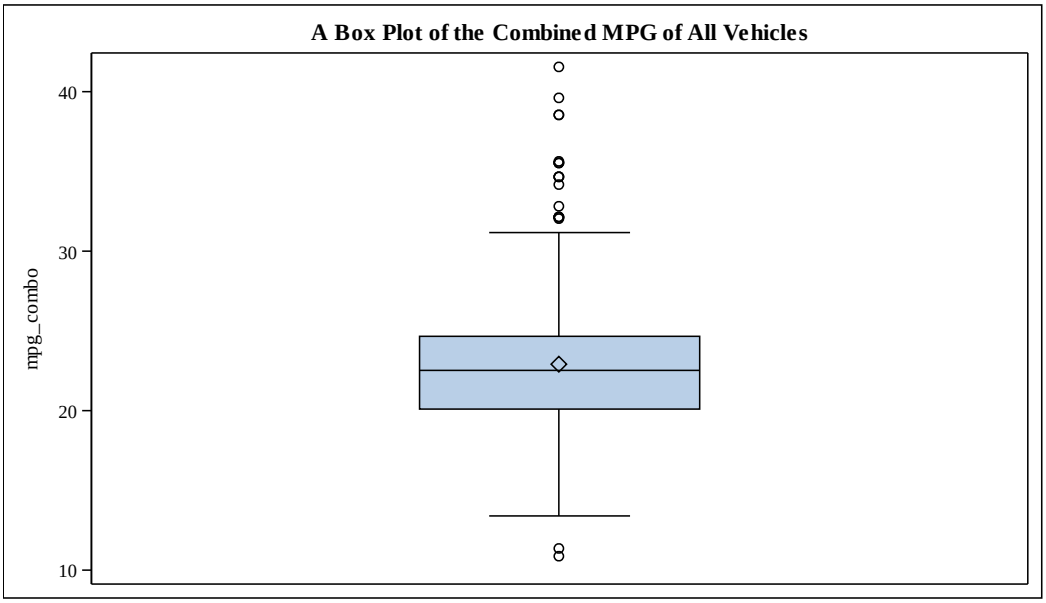
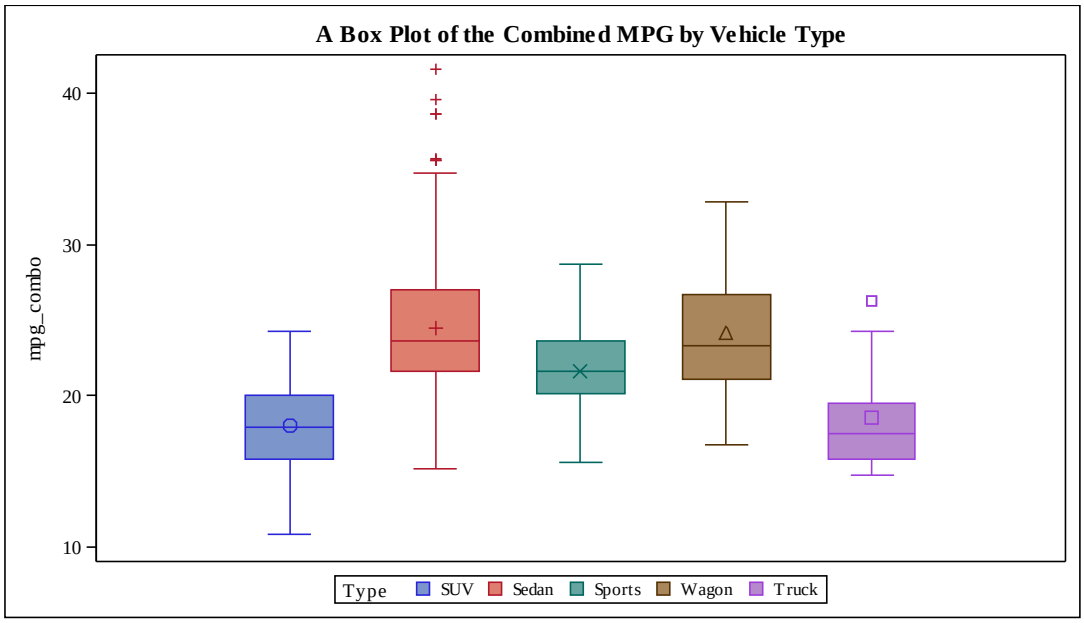
HW 1 Spring 2018

Problem 1 a



1a) The box plot of the combined MPG has several outliers above and below the quartiles according to the 1.5*IQR rule. Ignoring those outliers, the mean and median are fairly close together (roughly 22 combined miles per gallon) and the spread of the distribution is not wide, rather low variability. Most vehicles get between 20 and 25 miles per gallon (combined).

Problem 1 b



1b) When viewing the distribution of combined MPG separated by vehicle type, we see some interesting things. The combined MPG of SUVs and sports cars appear to be most like normal distributions - seemingly symmetric, with not wide spreads. Sedans' combined MPG have lots of variability and the most outliers. Combined MPG for wagons also have a wide spread but may not be symmetric. Trucks appear to be least fuel efficient among the vehicle types. The distribution of combined MPG of trucks is right skewed and has at least one outlier.

Problem 1 c

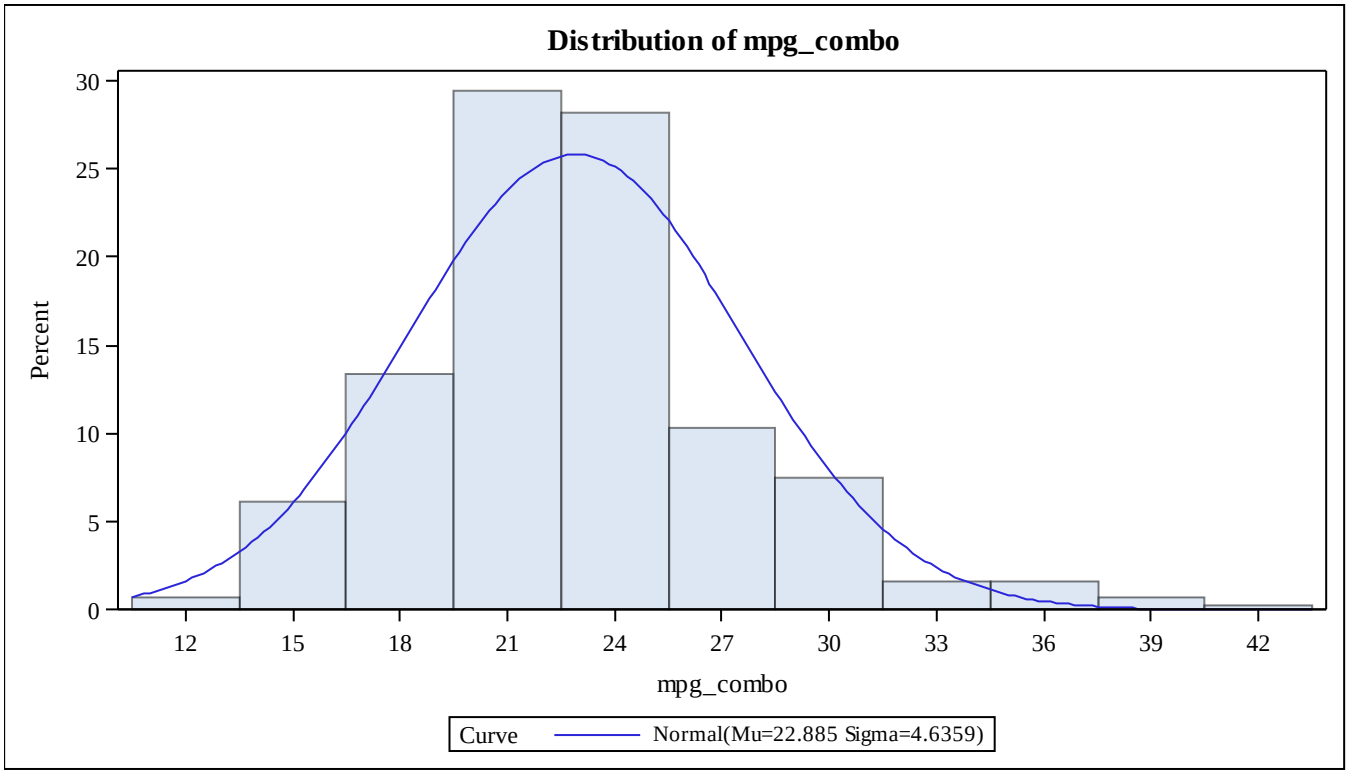
Variable:
mpg_combo

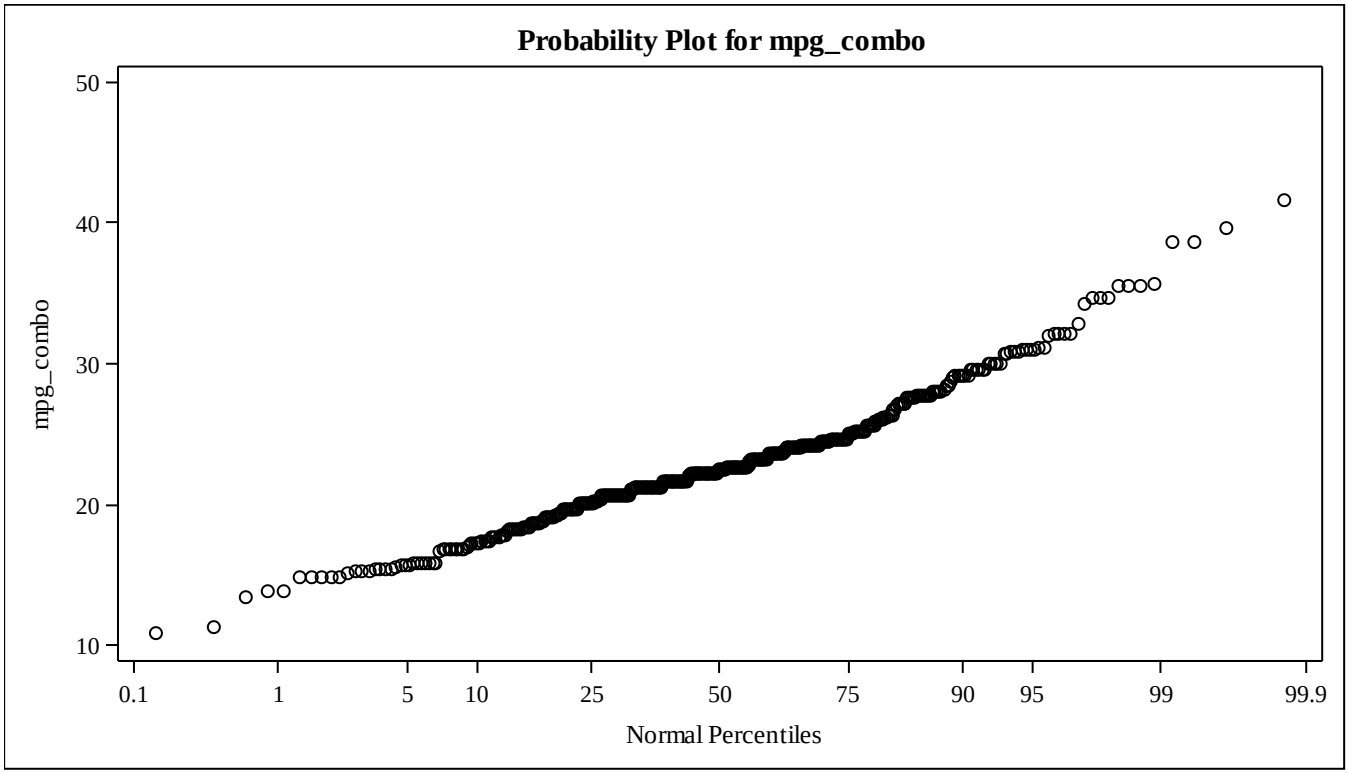
Moments			
N	425	**Sum Weights**	425
Mean	22.8847059	**Sum Observations**	9726
Std Deviation	4.6358734	**Variance**	21.4913221
Skewness	0.71081	**Kurtosis**	1.25934384
Uncorrected SS	231688.97	**Corrected SS**	9112.32059
Coeff Variation	20.2575179	**Std Error Mean**	0.22487289

Basic Statistical Measures			
Location		**Variability**	
Mean	22.88471	**Std Deviation**	4.63587
Median	22.50000	**Variance**	21.49132
Mode	21.15000	**Range**	30.70000
		Interquartile Range	4.55000

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.967372	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.103925	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.696027	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	3.776247	**Pr > A-Sq**	<0.0050

Extreme Observations					
Lowest			**Highest**		
Value	**Model**	**Obs**	**Value**	**Model**	**Obs**
10.90	H2	165	35.7	Echo 2dr auto	381
11.35	Excursion 6.8 XLT	119	38.6	Echo 2dr manual	380
13.45	G500	250	38.6	Echo 4dr	382
13.80	Discovery SE	215	39.6	Civic HX 2dr	154
13.80	Range Rover HSE	214	41.6	Jetta GLS TDI 4dr	402





1c) The mean, median, and mode of MPG_Combo are similar in value which would suggest symmetry. However, MPG_Combo has relatively low variance (21.22) and a small IQR (4.55) with several outliers since the vehicles are of various type. Some extremely fuel-inefficient vehicles are the SUVs: Hummer H2, Ford Excursion 6.8 XLT, Mercedes-Benz G500, and Land Rover's Discovery SE and Range Rover HSE. Some extremely fuel-efficient vehicles are: three versions of the Toyota Echo, the Honda Civic, and the Volkswagen Jetta GLS TDI. The normality test fails and suggests that the MPG_Combo distribution is not normal.

Variable:
Invoice

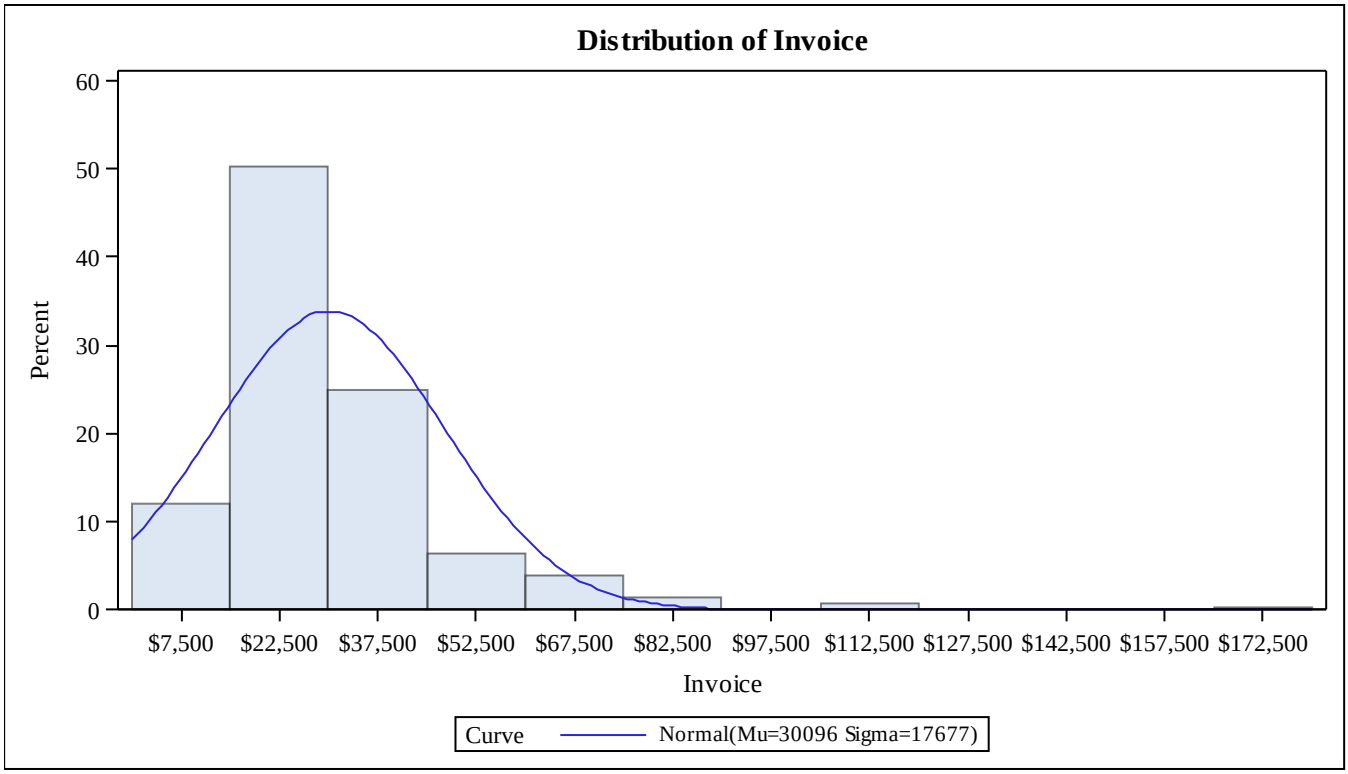
Moments			
N	425	**Sum Weights**	425
Mean	30096.48	**Sum Observations**	12791004
Std Deviation	17677.3562	**Variance**	312488924
Skewness	2.82591763	**Kurtosis**	13.8776543
Uncorrected SS	5.17459E11	**Corrected SS**	1.32495E11
Coeff Variation	58.735627	**Std Error Mean**	857.477729

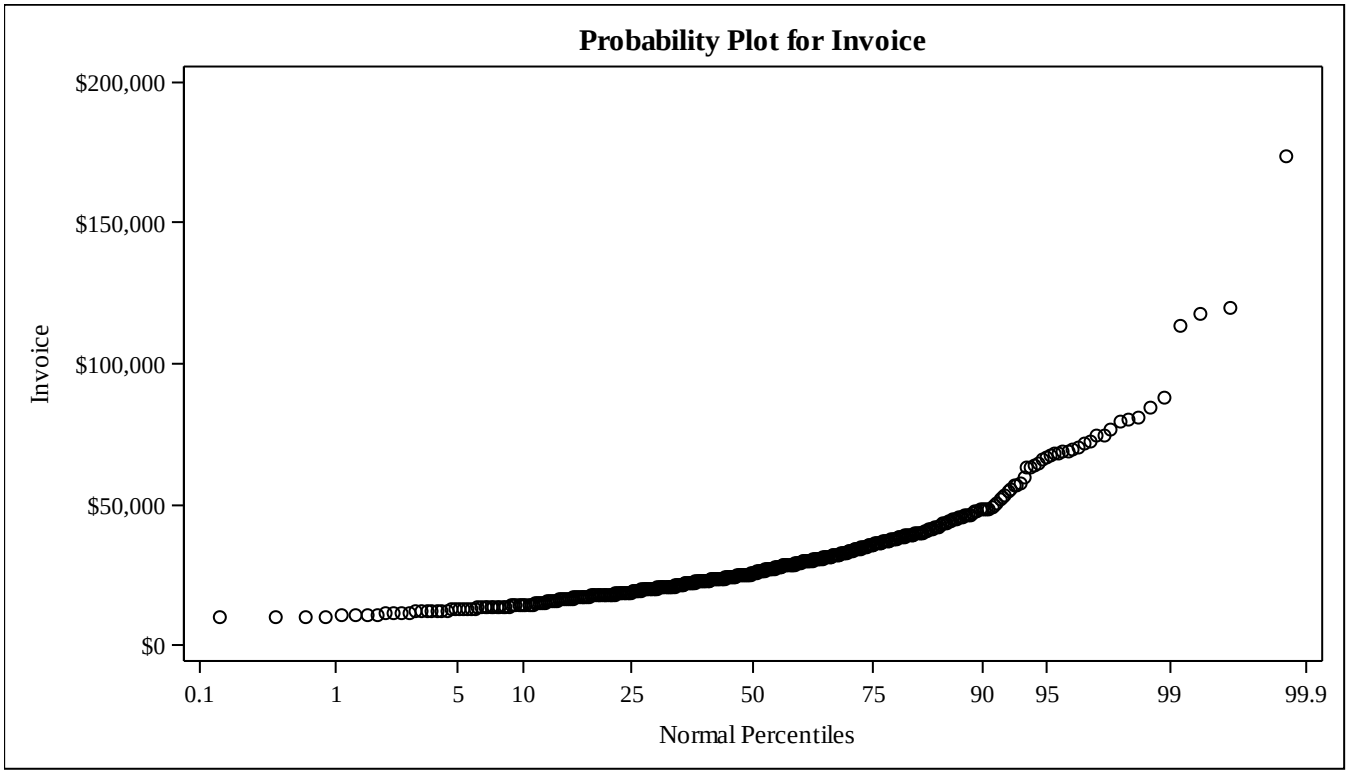
Basic Statistical Measures			
Location		**Variability**	
Mean	30096.48	**Std Deviation**	17677
Median	25672.00	**Variance**	312488924
Mode	14207.00	**Range**	163685
		Interquartile Range	16804

Note: The mode displayed is the smallest of 3 modes with a count of 2.

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

Extreme Observations					
Lowest			**Highest**		
Value	**Model**	**Obs**	**Value**	**Model**	**Obs**
9875	Rio 4dr manual	205	88324	CL500 2dr	260
10107	Accent 2dr hatch	167	113388	SL55 AMG 2dr	269
10144	Echo 2dr manual	380	117854	SL600 convertible 2dr	270
10319	Ion1 4dr	344	119600	CL600 2dr	261
10642	Echo 4dr	382	173560	911 GT2 2dr	333





The distribution of Invoice is not normal as visible from the probability plot and strongly skewed as visible from the histogram. The cheapest cars are Kia Rio, Hyundai Accent, two versions of the Toyota Echo, and the Saturn Ion. Some very expensive cars are all European vehicles including the Porsche 911 GT2 and four Mercedes-Benz vehicles - CL500, CL600, SL55 AMG, and SL 600. The mean, median, and mode are quite different from each other and clarifies the asymmetry of the invoice variable's distribution.

Problem 1 d

Variable:
mpg_combo

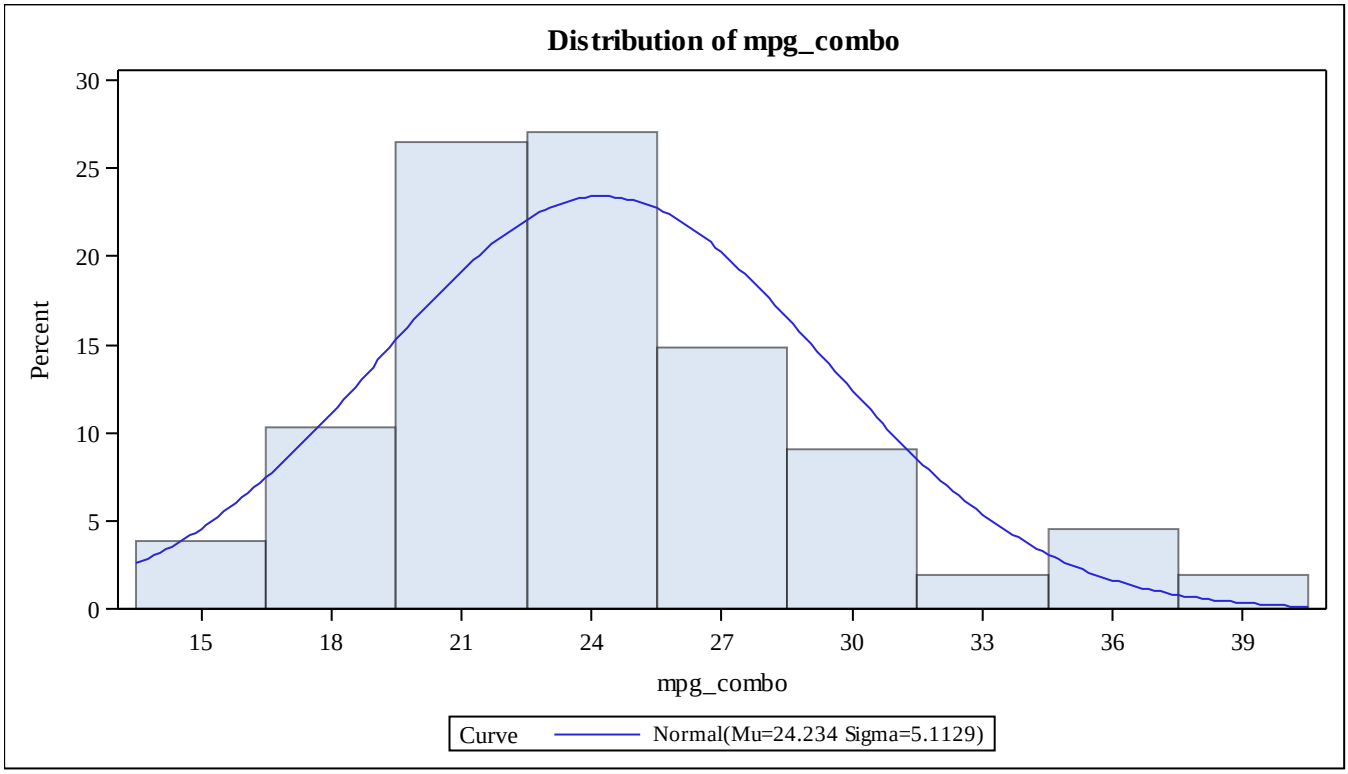
Origin=Asia

Moments			
N	155	**Sum Weights**	155
Mean	24.233871	**Sum Observations**	3756.25
Std Deviation	5.11289729	**Variance**	26.1417187
Skewness	0.71517885	**Kurtosis**	0.45333419
Uncorrected SS	95054.3025	**Corrected SS**	4025.82468
Coeff Variation	21.0981452	**Std Error Mean**	0.41067779

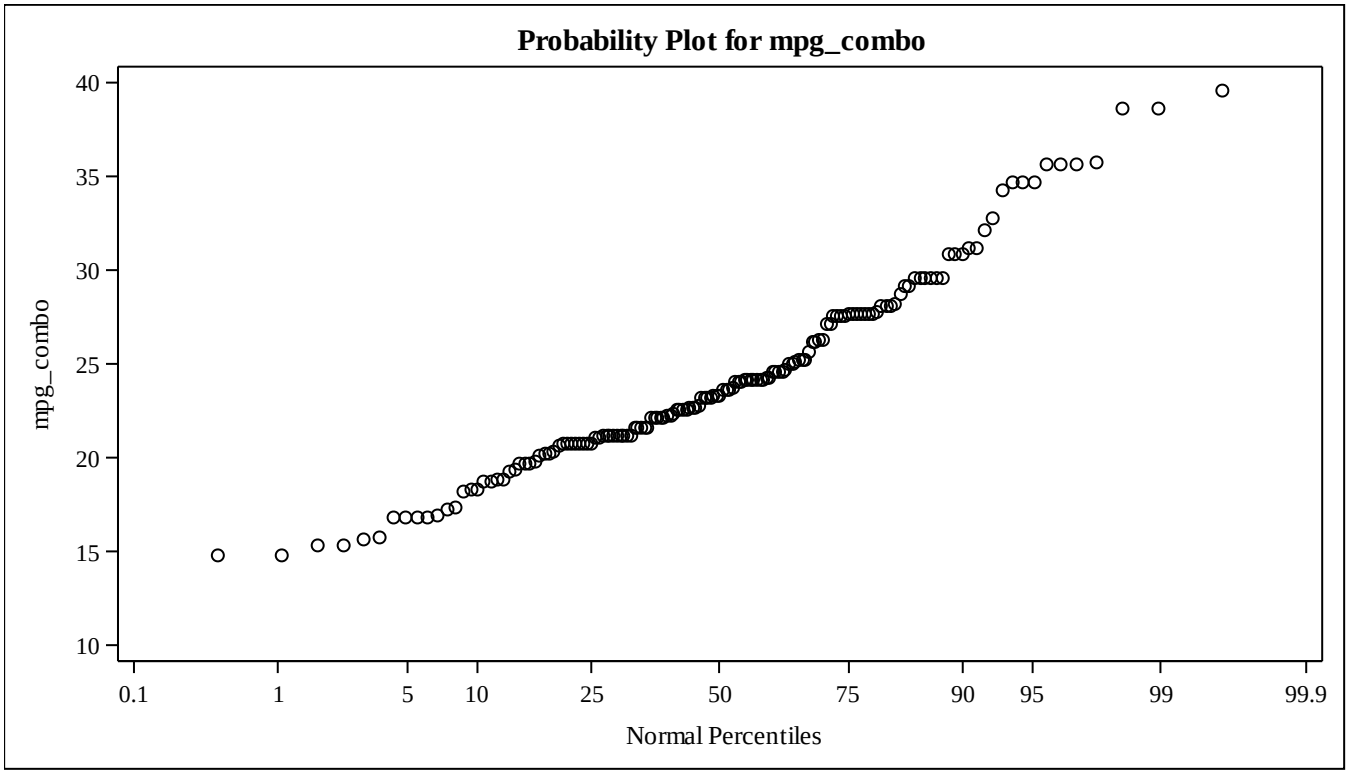
Basic Statistical Measures			
Location		**Variability**	
Mean	24.23387	**Std Deviation**	5.11290
Median	23.35000	**Variance**	26.14172
Mode	21.15000	**Range**	24.80000
		Interquartile Range	7.00000

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.959591	**Pr < W**	0.0002
Kolmogorov-Smirnov	**D**	0.105193	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.315702	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.837931	**Pr > A-Sq**	<0.0050

Origin=Asia



Origin=Asia



Variable:
Invoice

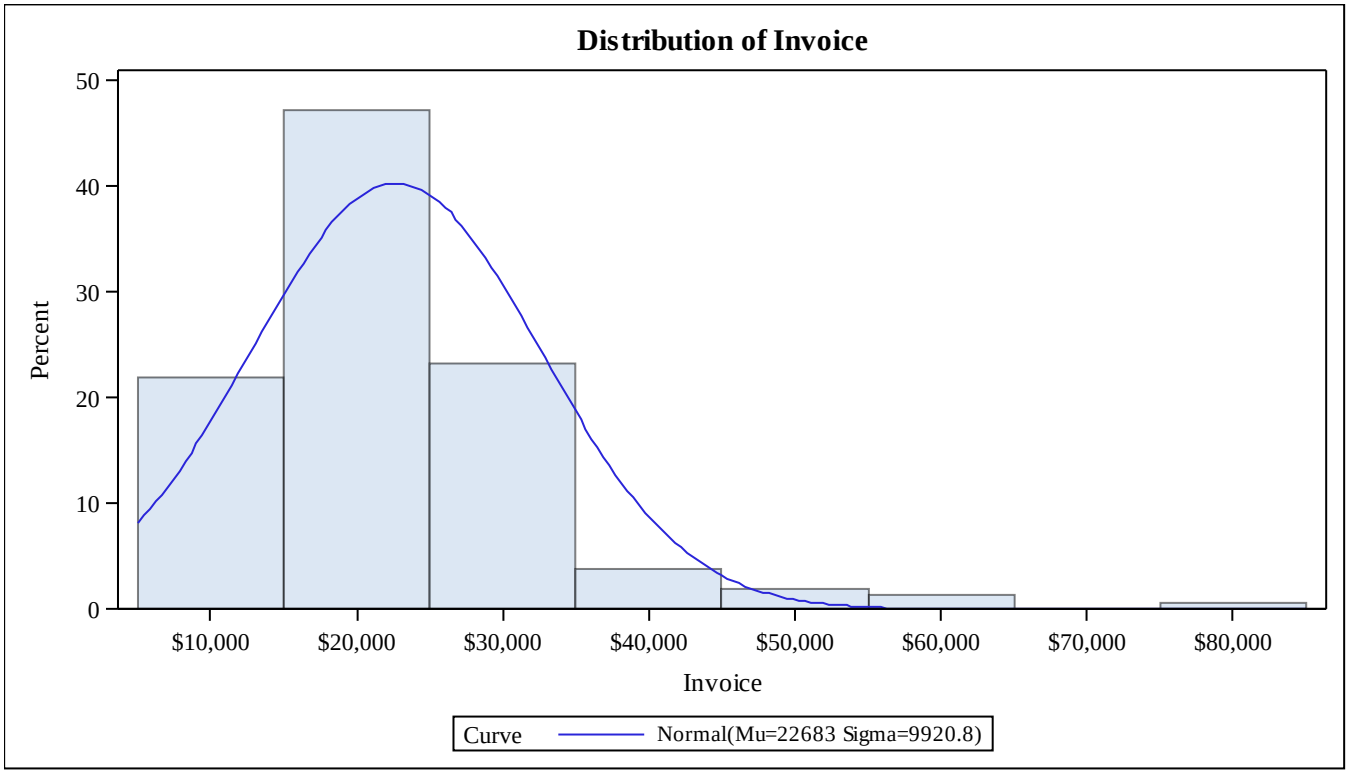
Origin=Asia

Moments			
N	155	**Sum Weights**	155
Mean	22682.9419	**Sum Observations**	3515856
Std Deviation	9920.81585	**Variance**	98422587.1
Skewness	2.08406116	**Kurtosis**	7.71656369
Uncorrected SS	9.4907E10	**Corrected SS**	1.51571E10
Coeff Variation	43.736901	**Std Error Mean**	796.859105

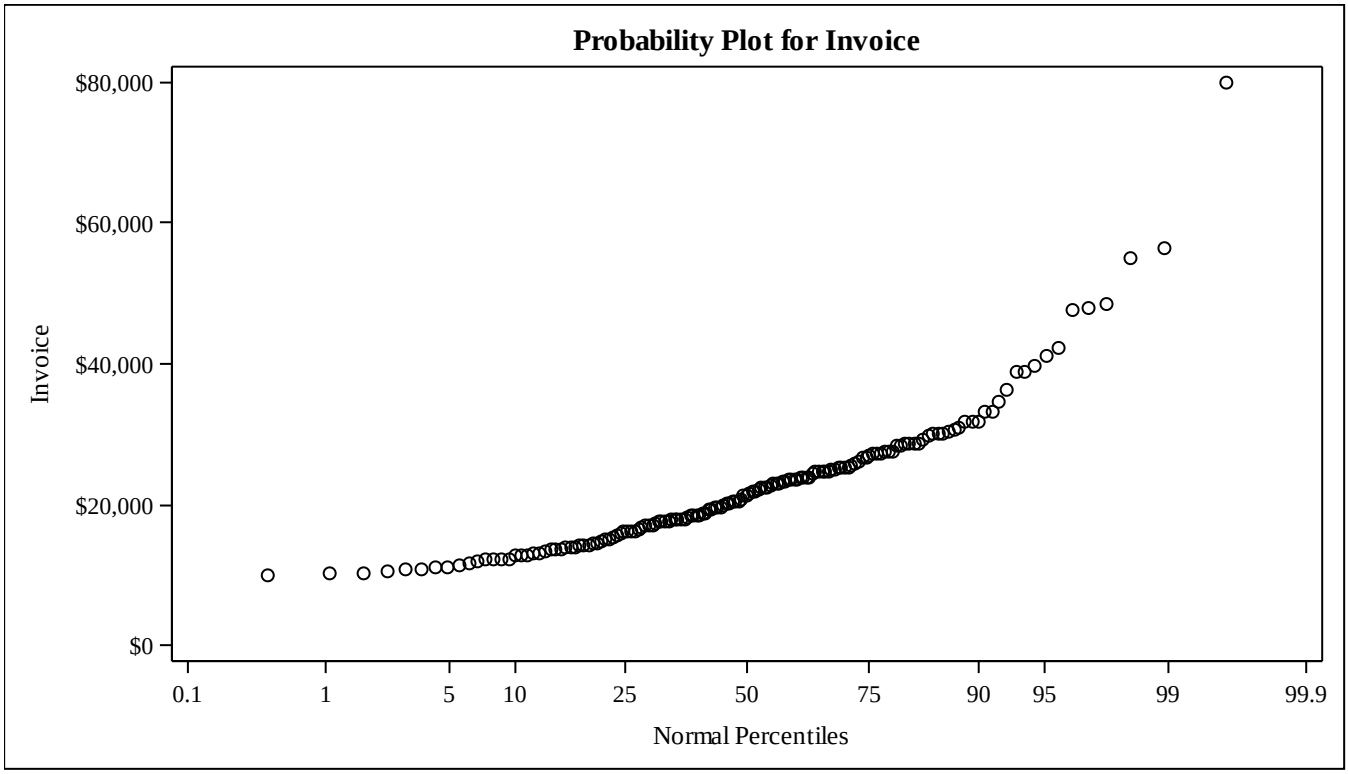
Basic Statistical Measures			
Location		**Variability**	
Mean	22682.94	**Std Deviation**	9921
Median	21428.00	**Variance**	98422587
Mode	14207.00	**Range**	70103
		Interquartile Range	10770

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

Origin=Asia



Origin=Asia



Variable:
mpg_combo

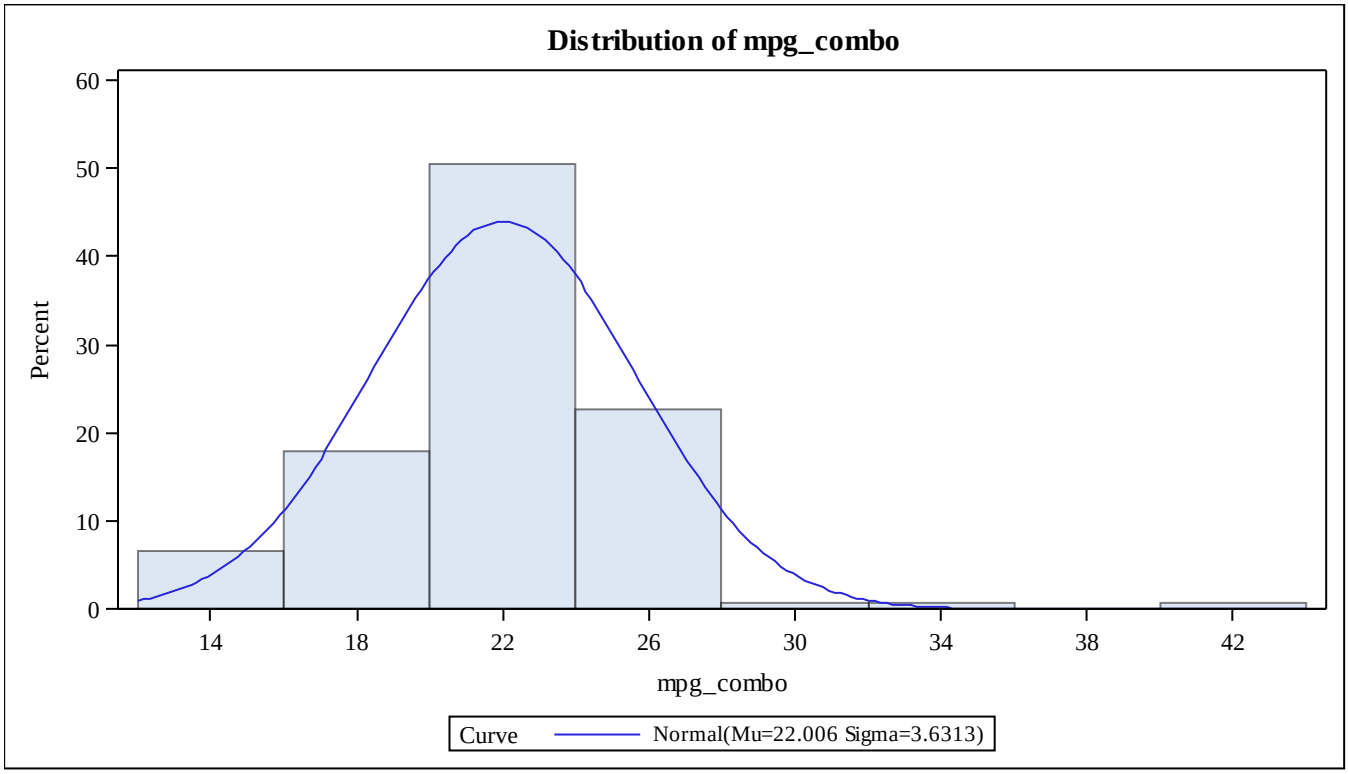
Origin=Europe

Moments			
N	123	**Sum Weights**	123
Mean	22.0060976	**Sum Observations**	2706.75
Std Deviation	3.63125427	**Variance**	13.1860076
Skewness	1.04311557	**Kurtosis**	6.47845029
Uncorrected SS	61173.6975	**Corrected SS**	1608.69293
Coeff Variation	16.5011278	**Std Error Mean**	0.32741917

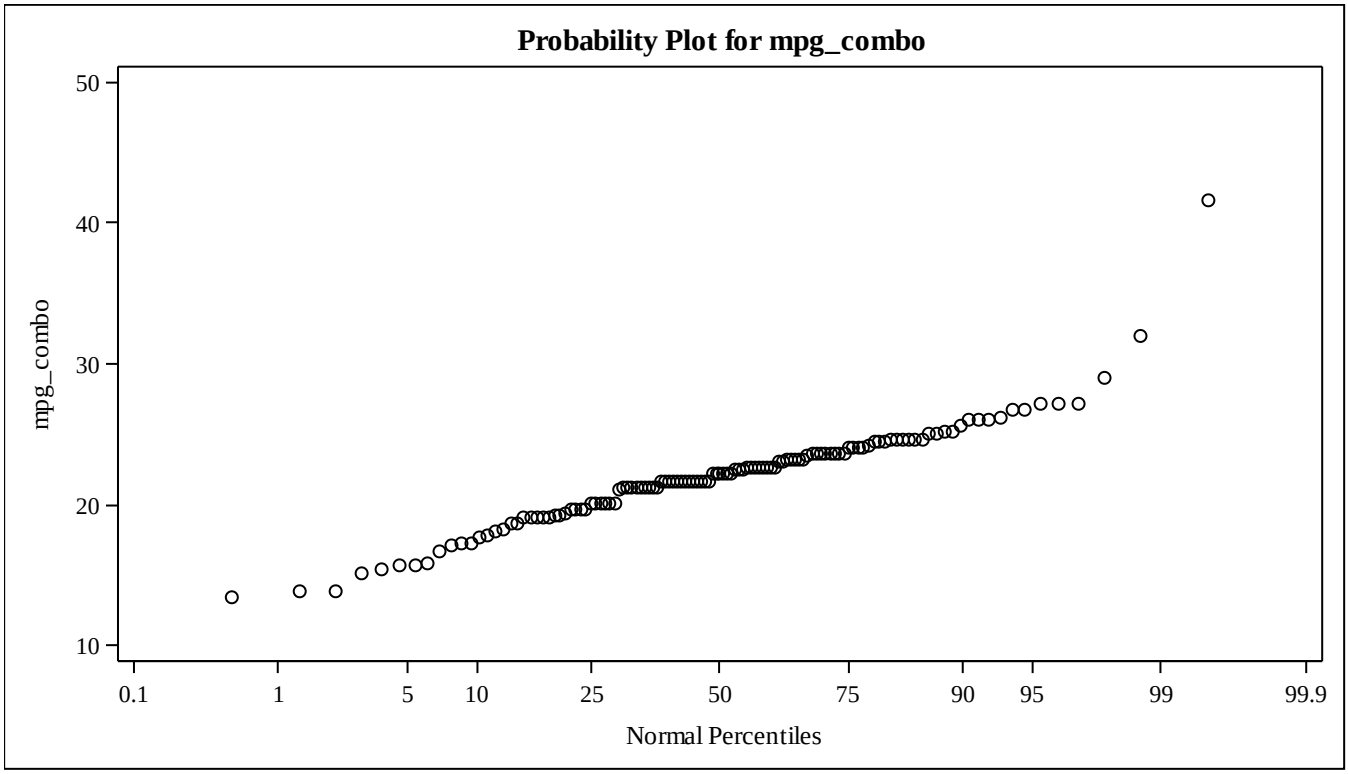
Basic Statistical Measures			
Location		**Variability**	
Mean	22.00610	**Std Deviation**	3.63125
Median	22.15000	**Variance**	13.18601
Mode	21.60000	**Range**	28.15000
		Interquartile Range	3.90000

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.915292	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.105997	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.247277	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.56949	**Pr > A-Sq**	<0.0050

Origin=Europe



Origin=Europe



Variable:
Invoice

Origin=Europe

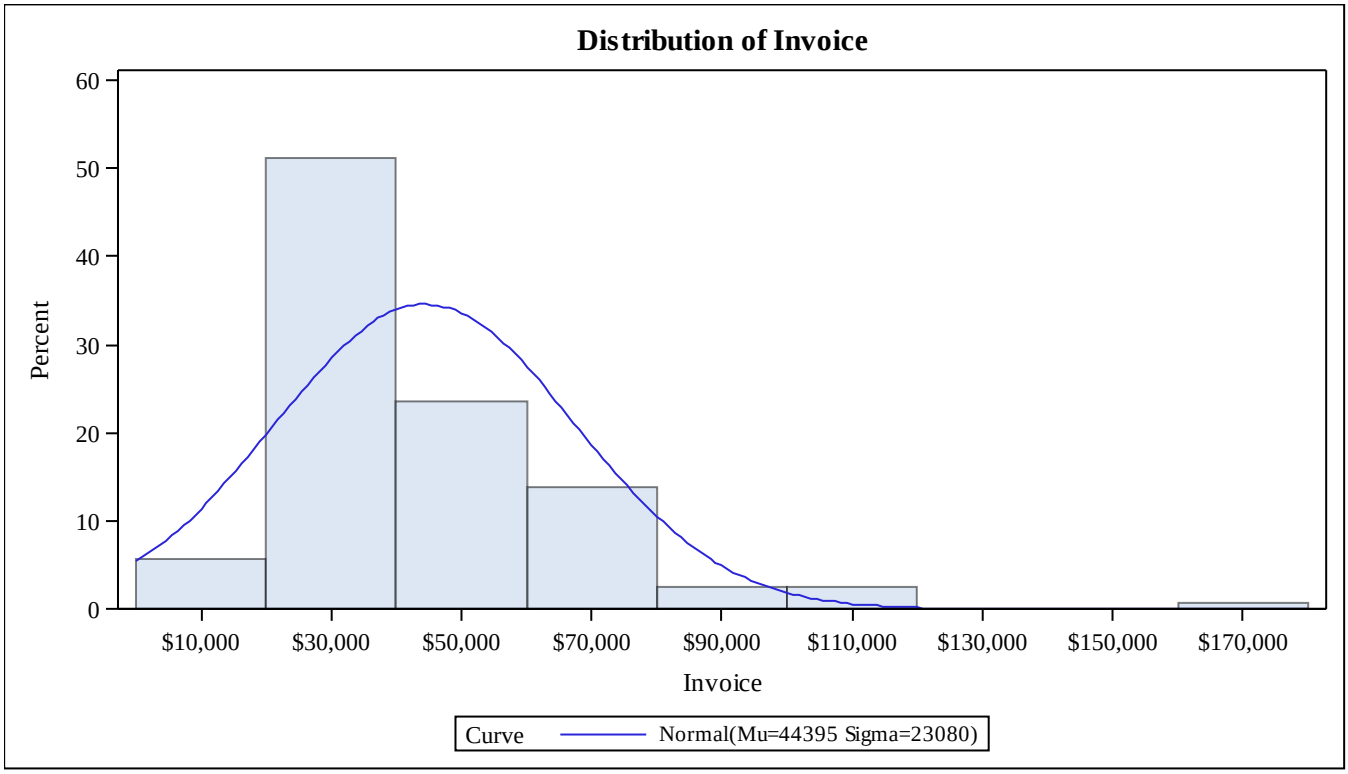
Moments			
N	123	**Sum Weights**	123
Mean	44395.0813	**Sum Observations**	5460595
Std Deviation	23080.3689	**Variance**	532703428
Skewness	2.36171691	**Kurtosis**	8.5886896
Uncorrected SS	3.07413E11	**Corrected SS**	6.49898E10
Coeff Variation	51.9885722	**Std Error Mean**	2081.08678

Basic Statistical Measures			
Location		**Variability**	
Mean	44395.08	**Std Deviation**	23080
Median	37575.00	**Variance**	532703428
Mode	19638.00	**Range**	158123
		Interquartile Range	20628

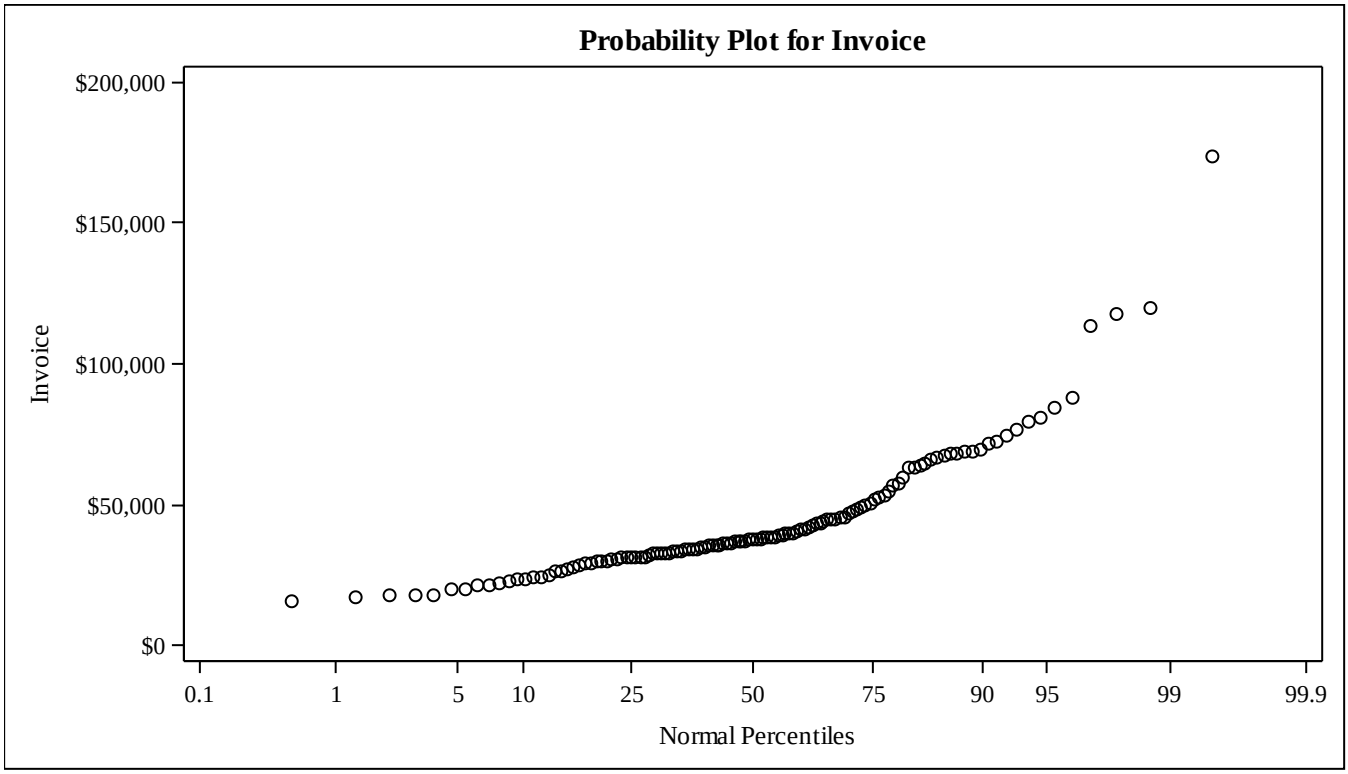
Note: The mode displayed is the smallest of 2 modes with a count of 2.

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

Origin=Europe



Origin=Europe



Variable:
mpg_combo

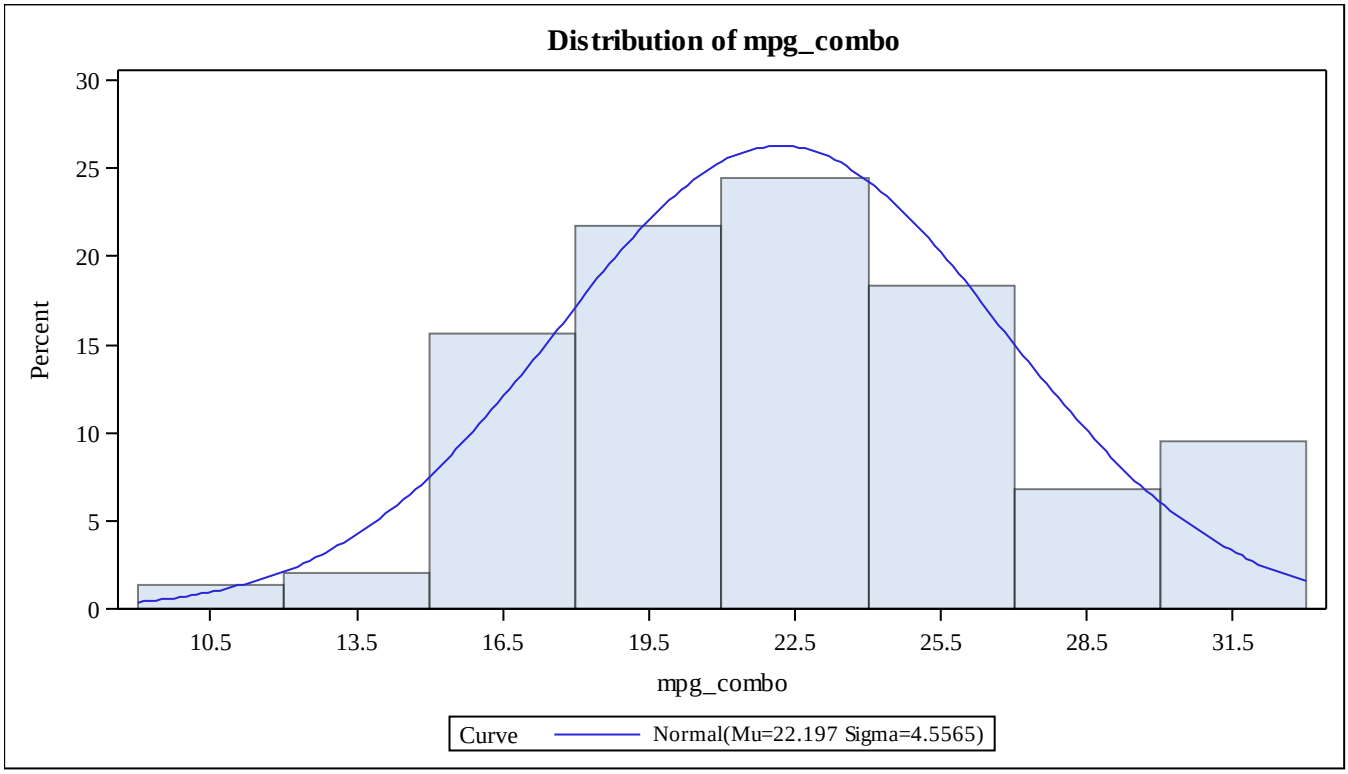
Origin=USA

Moments			
N	147	**Sum Weights**	147
Mean	22.1972789	**Sum Observations**	3263
Std Deviation	4.55653145	**Variance**	20.7619788
Skewness	0.2428057	**Kurtosis**	-0.2619995
Uncorrected SS	75460.97	**Corrected SS**	3031.24891
Coeff Variation	20.5274325	**Std Error Mean**	0.37581638

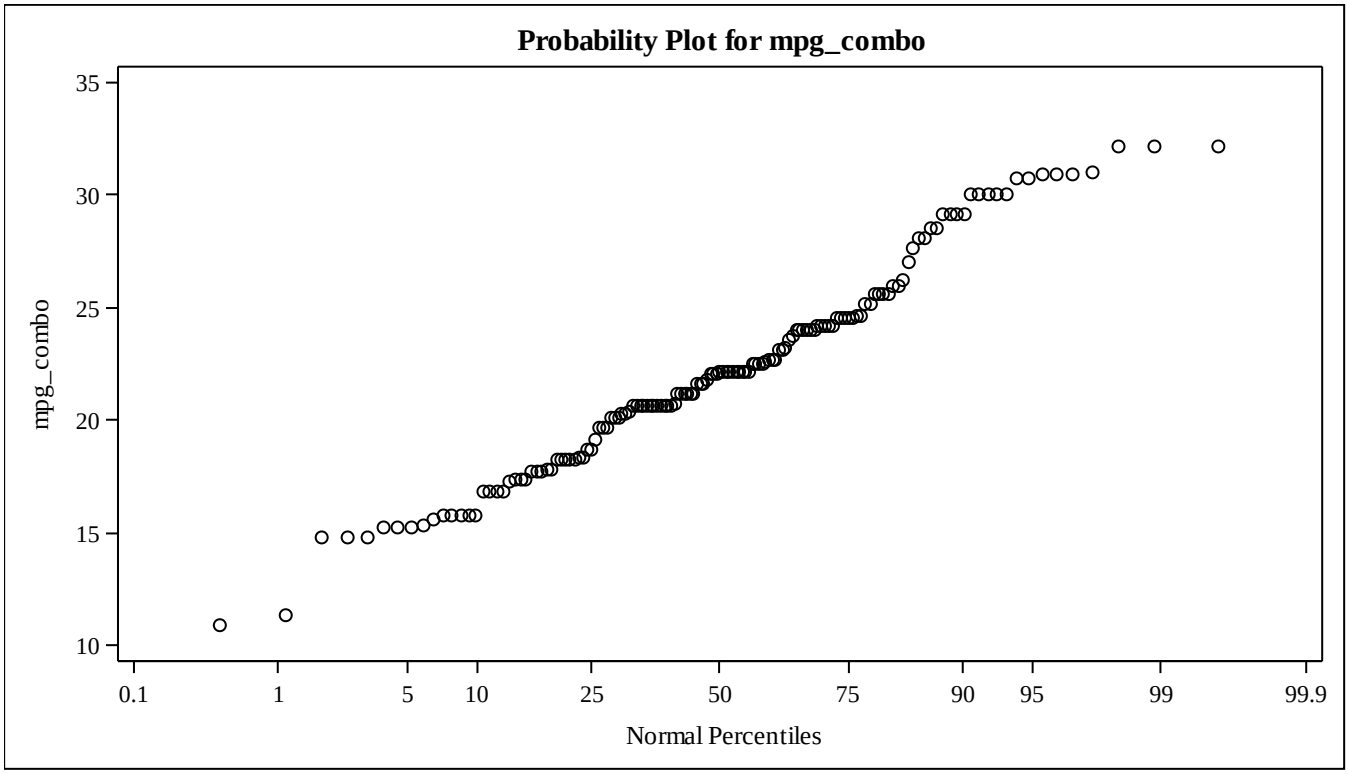
Basic Statistical Measures			
Location		**Variability**	
Mean	22.19728	**Std Deviation**	4.55653
Median	22.15000	**Variance**	20.76198
Mode	20.60000	**Range**	21.25000
		Interquartile Range	5.80000

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.97656	**Pr < W**	0.0128
Kolmogorov-Smirnov	**D**	0.075122	**Pr > D**	0.0415
Cramer-von Mises	**W-Sq**	0.152181	**Pr > W-Sq**	0.0227
Anderson-Darling	**A-Sq**	1.07095	**Pr > A-Sq**	0.0083

Origin=USA



Origin=USA



Variable:
Invoice

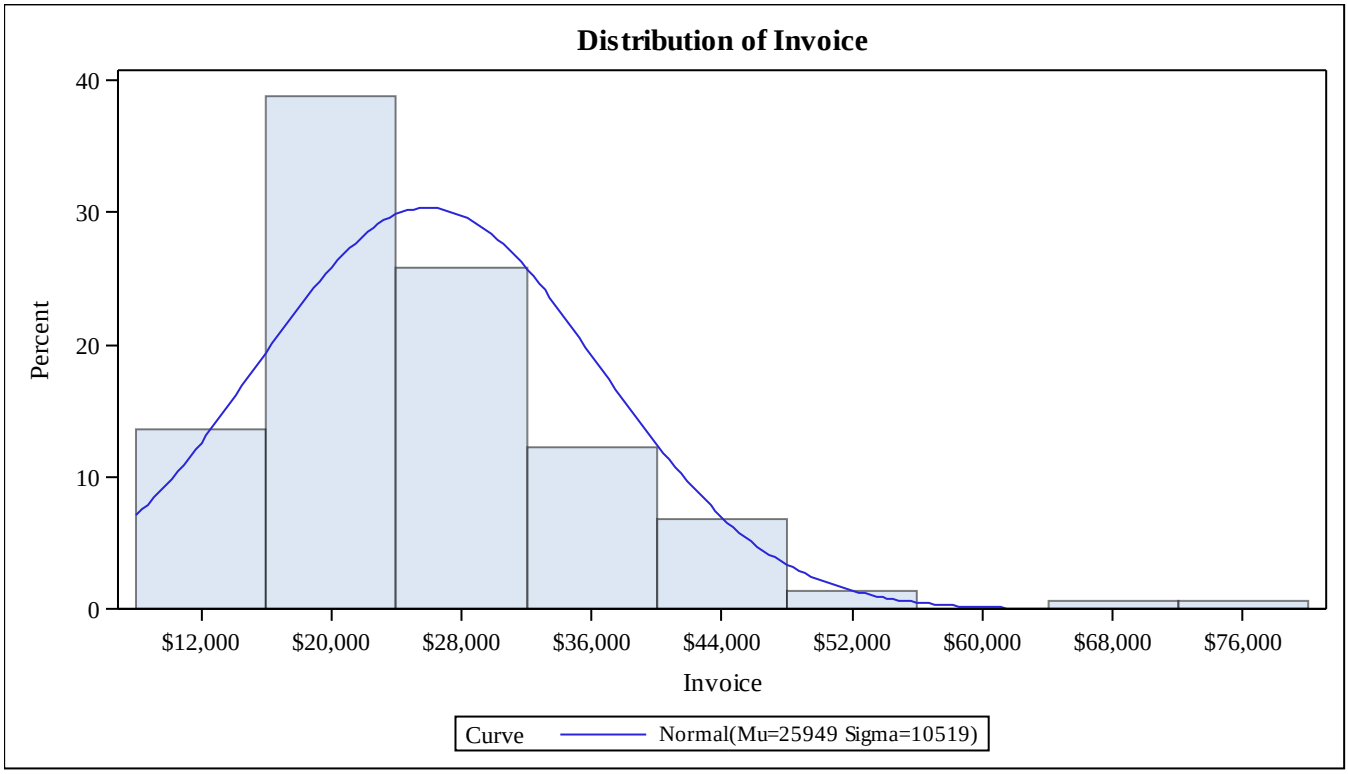
Origin=USA

Moments			
N	147	**Sum Weights**	147
Mean	25949.3401	**Sum Observations**	3814553
Std Deviation	10518.7222	**Variance**	110643517
Skewness	1.5375037	**Kurtosis**	3.95734687
Uncorrected SS	1.15139E11	**Corrected SS**	1.6154E10
Coeff Variation	40.5356057	**Std Error Mean**	867.569584

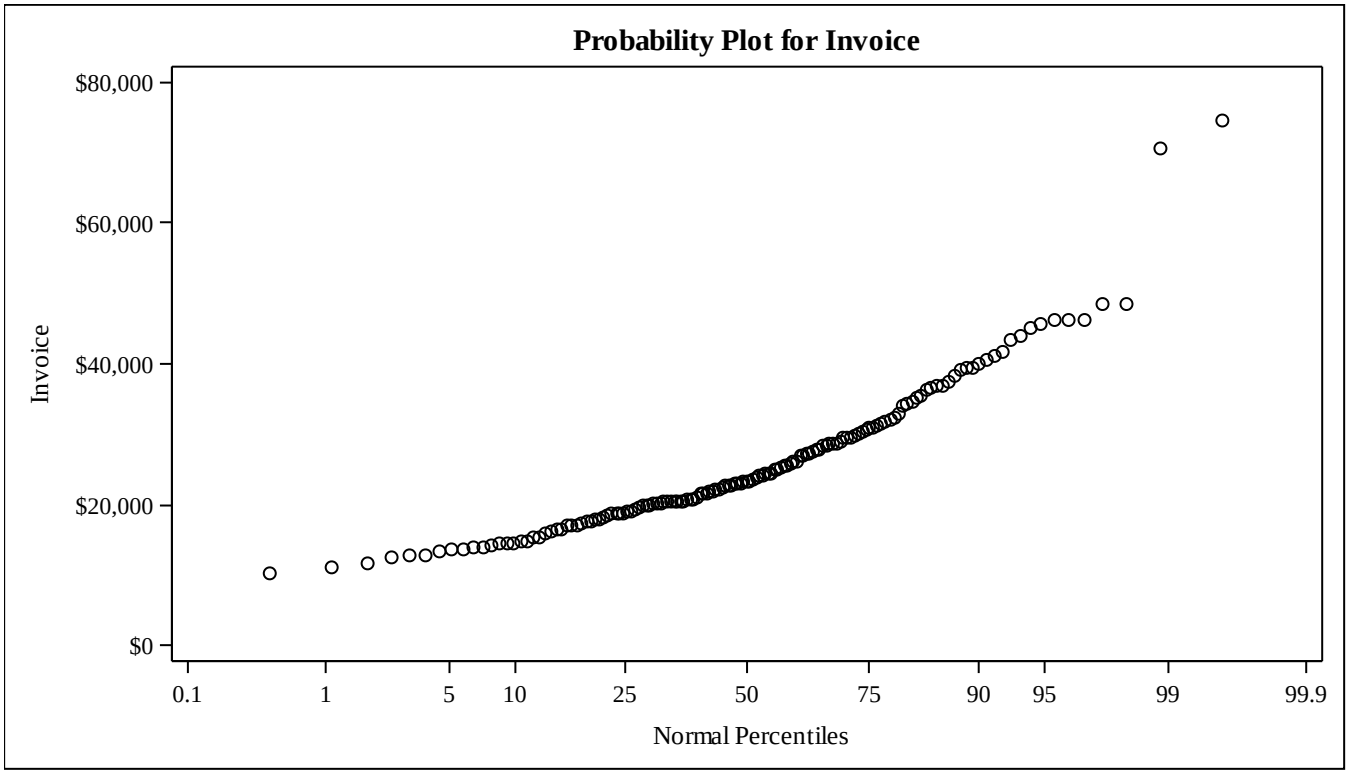
Basic Statistical Measures			
Location		**Variability**	
Mean	25949.34	**Std Deviation**	10519
Median	23217.00	**Variance**	110643517
Mode	.	**Range**	64132
		Interquartile Range	11965

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

Origin=USA



Origin=USA



1d) There are 155 cars originating from Asia, 123 European cars, and 147 originating from USA in the data set. None of the distributions of MPG_Combo or Invoice by Origin are normal according to the tests of normality, histograms, and probability plots. The distribution of Invoice (in dollars), is skewed and asymmetric for each of the 3 continents. The respective mean, median, and mode for each of the 3 distributions are quite different from each other. For MPG_Combo, the American and European cars have similar means (22) and medians (22.15) with low standard deviations (3.63 and 4.56 respectively).

Problem 2 a

Variable:
Invoice

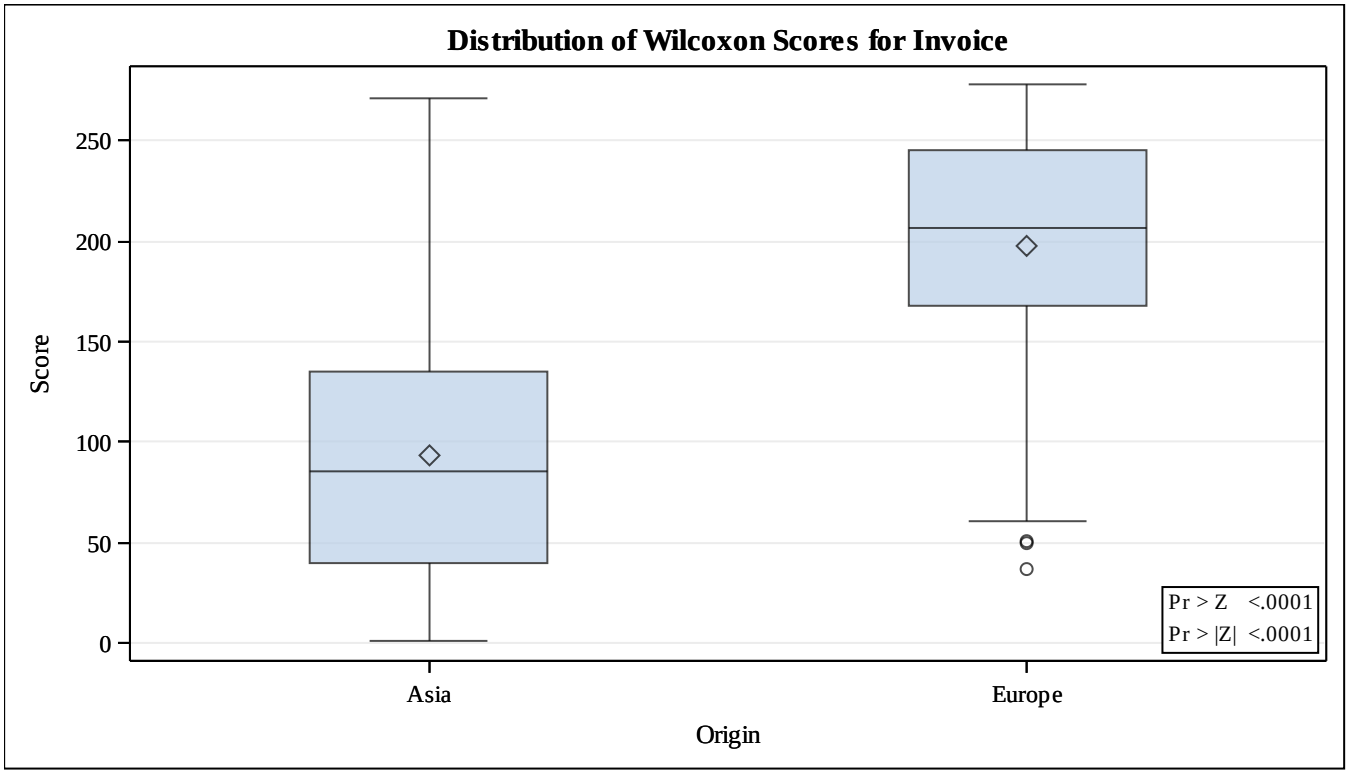
Tests for Location: Mu0=22000				
Test	**Statistic**		**p Value**	
Student's t	**t**	9.442204	**Pr > |t|**	<.0001
Sign	**M**	58.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	21930	**Pr >= |S|**	<.0001

2a) Since Invoice is not normal and not symmetric, we use the sign test for the null value of $22000. Based on the test for location table of proc univariate, we strongly reject the null hypothesis in favor of the alternative that the median invoice price of all vehicles is not $22000.

Problem 2 b

Wilcoxon Scores (Rank Sums) for Variable Invoice Classified by Variable Origin					
Origin	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Asia	155	14434.0	21622.50	665.778401	93.122581
Europe	123	24347.0	17158.50	665.778401	197.943089
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	24347.0000
Normal Approximation	
Z	10.7964
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	



2b) The distributions of Invoice for Europe and Asia are not normal from the normality tests above. Comparing the two distributions of the Invoice between European and Asian cars, we see the Wilcoxon rank sum test rejects the null that the two distributions are the same in favor of the alternative that the European cars tend to have more expensive Invoice prices. The box plots yield similar conclusions as the hypothesis test that the European cars have higher invoices.

Problem 3 a

Pearson Correlation Coefficients, N = 425 Prob > |r| under H0: Rho=0					
	Invoice	**Horsepower**	**Wheelbase**	**Length**	**mpg_combo**
Invoice	1.00000	0.82581 <.0001	0.14515 0.0027	0.16206 0.0008	-0.50457 <.0001
Horsepower	0.82581 <.0001	1.00000	0.38199 <.0001	0.37198 <.0001	-0.70511 <.0001
Wheelbase Wheelbase (IN)	0.14515 0.0027	0.38199 <.0001	1.00000	0.88882 <.0001	-0.56758 <.0001
Length Length (IN)	0.16206 0.0008	0.37198 <.0001	0.88882 <.0001	1.00000	-0.50977 <.0001
mpg_combo	-0.50457 <.0001	-0.70511 <.0001	-0.56758 <.0001	-0.50977 <.0001	1.00000

3a) Length and Wheelbase have strongest correlation among the 5 variables at 0.89 indicating as wheelbase increases the length of the vehicle tends to increase as well and vice versa. Invoice and Horsepower have a strong correlation of 0.82 indicating as Invoice price tends to increase so does the Horsepower of the car and vice versa. Horsepower and Combined MPG have a moderate negative correlation of -0.71 indicating as the Combined MPG tends to increase, the Horsepower tends to decrease and vice versa. The remaining pairwise correlations are much lower and indicate weaker linear relationships. The correlation hypothesis tests for each pair of variables are statistically significant as being unequal to 0 correlation. However, as aforementioned, there are only 3 pairs of moderately strong correlations.

Problem 3 b

Type=SUV

Pearson Correlation Coefficients, N = 60 Prob > |r| under H0: Rho=0					
	Invoice	**Horsepower**	**Wheelbase**	**Length**	**mpg_combo**
Invoice	1.00000	0.76589 <.0001	0.44030 0.0004	0.45230 0.0003	-0.73206 <.0001
Horsepower	0.76589 <.0001	1.00000	0.70261 <.0001	0.69163 <.0001	-0.75954 <.0001
Wheelbase Wheelbase (IN)	0.44030 0.0004	0.70261 <.0001	1.00000	0.93951 <.0001	-0.62128 <.0001
Length Length (IN)	0.45230 0.0003	0.69163 <.0001	0.93951 <.0001	1.00000	-0.59980 <.0001
mpg_combo	-0.73206 <.0001	-0.75954 <.0001	-0.62128 <.0001	-0.59980 <.0001	1.00000

Type=Sedan

Pearson Correlation Coefficients, N = 262 Prob > |r| under H0: Rho=0					
	Invoice	**Horsepower**	**Wheelbase**	**Length**	**mpg_combo**
Invoice	1.00000	0.85442 <.0001	0.59361 <.0001	0.42909 <.0001	-0.63038 <.0001
Horsepower	0.85442 <.0001	1.00000	0.66364 <.0001	0.55883 <.0001	-0.75485 <.0001
Wheelbase Wheelbase (IN)	0.59361 <.0001	0.66364 <.0001	1.00000	0.85600 <.0001	-0.69395 <.0001
Length Length (IN)	0.42909 <.0001	0.55883 <.0001	0.85600 <.0001	1.00000	-0.60801 <.0001
mpg_combo	-0.63038 <.0001	-0.75485 <.0001	-0.69395 <.0001	-0.60801 <.0001	1.00000

Type=Sports

Pearson Correlation Coefficients, N = 49 Prob > |r| under H0: Rho=0					
	Invoice	**Horsepower**	**Wheelbase**	**Length**	**mpg_combo**
Invoice	1.00000	0.80000 <.0001	-0.05278 0.7187	0.34370 0.0156	-0.59893 <.0001
Horsepower	0.80000 <.0001	1.00000	0.31025 0.0300	0.56576 <.0001	-0.84301 <.0001
Wheelbase Wheelbase (IN)	-0.05278 0.7187	0.31025 0.0300	1.00000	0.67539 <.0001	-0.39329 0.0052
Length Length (IN)	0.34370 0.0156	0.56576 <.0001	0.67539 <.0001	1.00000	-0.59452 <.0001
mpg_combo	-0.59893 <.0001	-0.84301 <.0001	-0.39329 0.0052	-0.59452 <.0001	1.00000

Type=Truck

Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0					
	Invoice	**Horsepower**	**Wheelbase**	**Length**	**mpg_combo**
Invoice	1.00000	0.84423 <.0001	0.48689 0.0158	0.49544 0.0138	-0.60713 0.0017
Horsepower	0.84423 <.0001	1.00000	0.74923 <.0001	0.71352 <.0001	-0.74895 <.0001
Wheelbase Wheelbase (IN)	0.48689 0.0158	0.74923 <.0001	1.00000	0.94404 <.0001	-0.75378 <.0001
Length Length (IN)	0.49544 0.0138	0.71352 <.0001	0.94404 <.0001	1.00000	-0.74183 <.0001
mpg_combo	-0.60713 0.0017	-0.74895 <.0001	-0.75378 <.0001	-0.74183 <.0001	1.00000

Type=Wagon

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0					
	Invoice	**Horsepower**	**Wheelbase**	**Length**	**mpg_combo**
Invoice	1.00000	0.83032 <.0001	0.60500 0.0004	0.47532 0.0079	-0.73759 <.0001
Horsepower	0.83032 <.0001	1.00000	0.64068 0.0001	0.49217 0.0057	-0.84074 <.0001
Wheelbase Wheelbase (IN)	0.60500 0.0004	0.64068 0.0001	1.00000	0.83381 <.0001	-0.66254 <.0001
Length Length (IN)	0.47532 0.0079	0.49217 0.0057	0.83381 <.0001	1.00000	-0.62878 0.0002
mpg_combo	-0.73759 <.0001	-0.84074 <.0001	-0.66254 <.0001	-0.62878 0.0002	1.00000

3b) There are 60 SUVs, and 242 sedans, 49 sports cars, 24 trucks, and 30 wagons in this cars data set. Among SUVs, there are relatively high, positive correlations between the following pairs of variables: Horsepower/Invoice (0.77), Horsepower/Wheelbase (0.70), Horsepower/Length (0.69), and Length/Wheelbase (0.94). Among sedans, there are relatively high, positive correlations between the following pairs of variables: Horsepower/Invoice (0.85) and Length/Wheelbase (0.86). Among sports cars, there are relatively high, positive correlations between Horsepower and Invoice (0.80). Among trucks, there are relatively high, positive correlations between the following pairs of variables: Horsepower/Invoice (0.84), Horsepower/Wheelbase (0.75), Horsepower/Length (0.71), and Length/Wheelbase (0.94). Among wagons, there are relatively high, positive correlations between the following pairs of variables: Horsepower/Invoice (0.83) and Length/Wheelbase (0.83). These relatively high, positive correlations indicate as one variable tends to increases so does the other variable.

Among SUVs, there are relatively high, negative correlations between the following pairs of variables: Horsepower/MPG_Combo (-0.76) and Invoice/MPG_Combo (-0.73). Among sedans, there are relatively high, negative correlations between the following pairs of variables: Horsepower/MPG_Combo (-0.75) and Wheelbase/MPG_Combo (-0.69). Among sports cars, there are relatively high, negative correlations between Horsepower and MPG_Combo (-0.84). Among trucks, there are relatively high, negative correlations between the following pairs of variables: Horsepower/MPG_Combo (-0.75), Wheelbase/MPG_Combo (-0.75), and Length/MPG_Combo (-0.74). Among wagons, there are relatively high, negative correlations between the following pairs of variables: Horsepower/MPG_Combo (-0.84) and Invoice/MPG_Combo (-0.74). These relatively high, negative correlations indicate as one variable tends to increase, the other variable tends to decrease and vice versa.

The results here are more specific than the correlation results from part 3a). We do see the pairs of variables Length/Wheelbase and Invoice/Horsepower from part a) as moderately strong positive correlations for all 5 vehicle types. We also see the moderately strong negative correlation of Horsepower/Combined MPG among all vehicle types which was the case in part 3a). The major difference between part 3b) and 3a) is the additional pairs of variables achieving moderately strong correlations - positive and negative. Horsepower/Length and Horsepower/Wheelbase are positively correlated for several vehicle types. Invoice/Combined MPG, Wheelbase/Combined MPG, and Length/Combined MPG are negatively correlated for several vehicle types.