Deliverable 6:

Number of Emails Being Sent by Age Group

Rebekah Sander

Research Question

Can the person's age group predict the number of emails sent each month?

Research Variables

<table>
<tr><th rowspan="2">WHO</th><th colspan="2">WHAT measurement is made on each</th><th rowspan="2">TYPE OF MEASURE</th></tr>
<tr><th>Name of Variable</th><th>Question Asked</th></tr>
<tr><td rowspan="2">A Person</td><td>Age Group</td><td>What age group is this person in?</td><td>Categorical Variable: levels =
Under 20;
20 to less than 40;
40 to less than 60;
60 and Older</td></tr>
<tr><td>Number of Emails</td><td>How many emails were sent by this person in one month?</td><td>Quantitative Variable
Unit: emails</td></tr>
<tr><td colspan="4">One quantitative variable being tested among one categorical variable with K levels.</td></tr>
<tr><td colspan="4">1. One-way ANOVA
2. Welch's Test on Raw Data
3. Kruskal Wallis
4. Welch's Test on Ranked Data</td></tr>
</table>

SAS Code: Examining the Data

```
/* Check for and fix miscoding/missing values */
PROC FREQ DATA=WORK.emails;
TABLE 'Age Group'n 'Number of Emails'n;
run;

Proc Contents data=work.emails varnum;
run;

/*Fixing number of emails*/
data work.emails;
    set work.emails;
    if 'Number of Emails'n = -1 or 'Number of Emails'n = '.'
        or 'Number of Emails'n = 99999 or 'Number of Emails'n = 'NA'
        or 'Number of Emails'n = null then delete;
    'Number of Emails num'n = input('Number of Emails'n, ?? best32.)
    drop 'Number of Emails'n null;
    rename 'Number of Emails num'n='Number of Emails'n;
run;
```

SAS Code: Examining the Data

```
/*fixing age group*/
*20_to_40, 40_to_60,less_then_20, less_than_20, over_60;
data work.emails;
set work.emails;
length 'Age Group'n $20;
    if 'Age Group'n = '20_to_40' then 'Age Group Category'n='B';
    else if 'Age Group'n = '40_to_60' then 'Age Group Category'n='C';
    else if 'Age Group'n = 'Less_then_20'
        or 'Age Group'n = 'Less_than_20' then 'Age Group Category'n='A';
    else if 'Age Group'n = 'Over_60' then 'Age Group Category'n='D';
drop 'Age Group'n
RUN;
```

```
Proc format;
Value $avgformat
'A'="Under 20"
'B'="20 to less than 40"
'C'="40 to less than 60"
'D'="60 and Older";
run;
```

```
data work.emails;
set work.emails;
format 'Age Group Category'n avgformat.;
rename 'Age Group Category'n='Age Group'n;
run;
```

SAS Code: Assessing Normality

```
/*QQ Plots and normality test*/
title 'Figures 1, 2, 3: QQ Plots for Number of Emails by Age Group';
proc univariate data=work.emails normaltest plots;
    var 'Number of Emails'n;
    class 'Age Group'n;
title;
```

```
PROC FREQ DATA=WORK.emails;
TABLE 'Age Group'n 'Number of Emails'n;
run;
```

```
Proc Contents data=work.emails varnum;
run;
```

```
Proc Means data = work.emails MAXDEC=2 n mean stddev median Qrange RANGE min Q1 Q3 max;
    var 'Number of Emails'n;
run;
```

Assessing Normality

- H_0: The data came from a population where the number of emails are normally distributed
- H_A: The data came from a population where the number of emails are not normally distributed.
- $\alpha = 0.10$

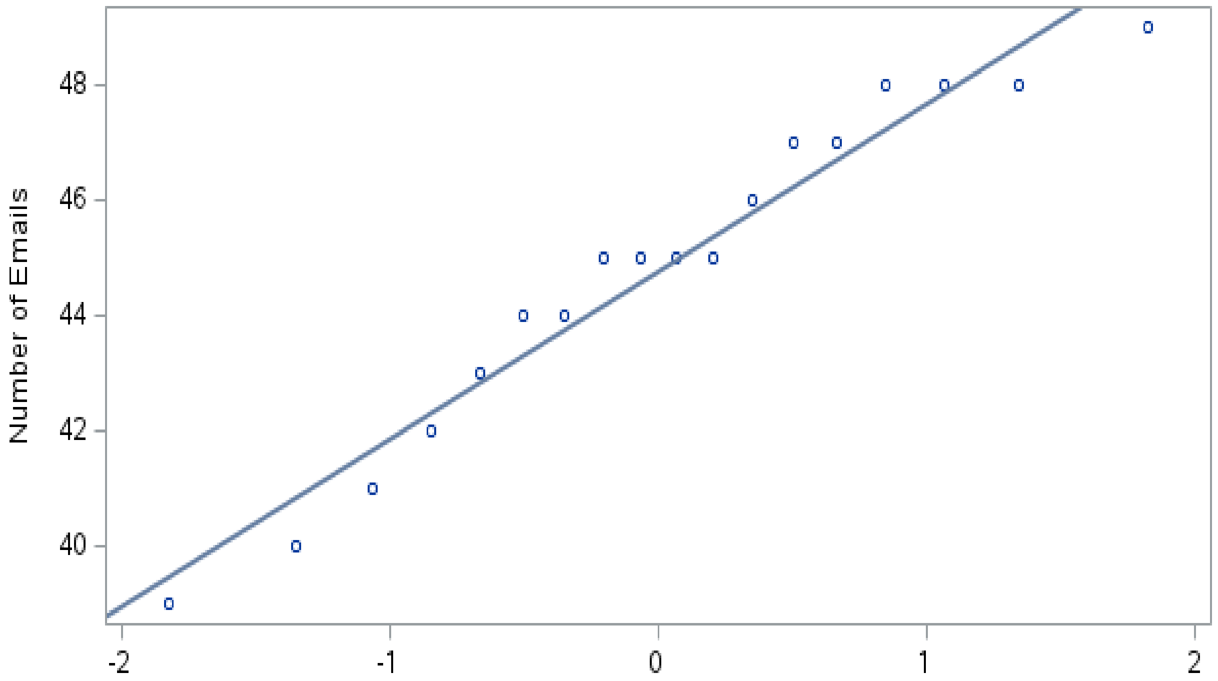
Age Group	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Under 20	18	29.03	18	29.03
20 to less than 40	14	22.58	32	51.61
40 to less than 60	16	25.81	48	77.42
60 and Older	14	22.58	62	100.00

- The sample sizes for all four categories are less than 30. Therefore, the $\bar{x}$ distribution cannot be assumed normal.

Assessing Normality—Under 20

- Normality Tests: All four of the tests for normality show p-values greater than $\alpha = 0.10$. Thus, we have evidence to suggest the x distribution is normal.
- QQ Plot: The data follows the agreement line with little deviation. This supports that the x distribution is normal.

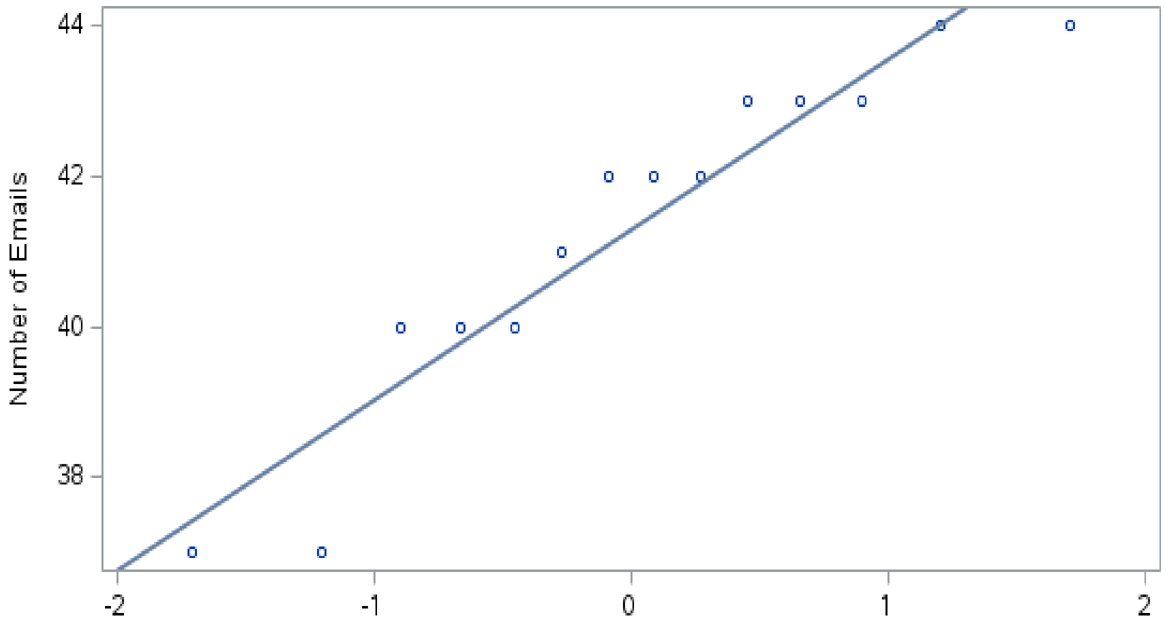
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.948626	Pr < W	0.4039
Kolmogorov-Smirnov	D	0.141636	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.053435	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.346307	Pr > A-Sq	>0.2500



Assessing Normality—20 to less than 40

- Normality Tests: Three of the tests for normality show p-values greater than $\alpha = 0.10$. Thus, we have evidence to suggest the x distribution is not normal.
- QQ Plot: The data follows the agreement line with fairly little deviation.

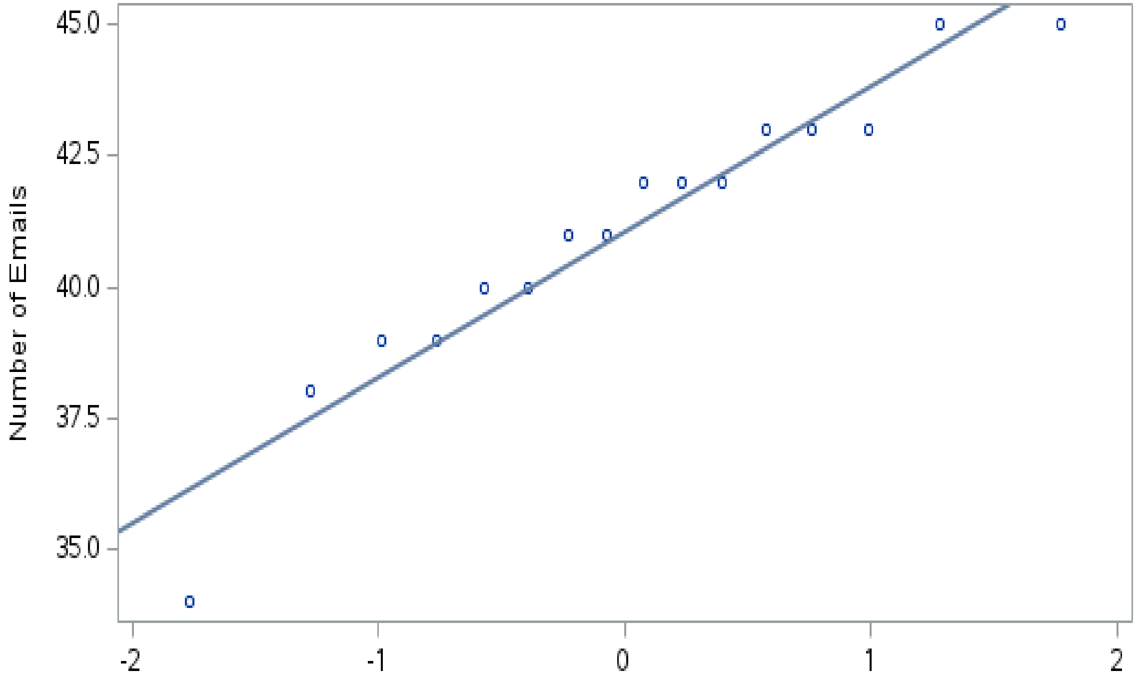
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.89198	Pr < W	0.0863
Kolmogorov-Smirnov	D	0.195037	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.081195	Pr > W-Sq	0.1906
Anderson-Darling	A-Sq	0.562125	Pr > A-Sq	0.1231



Assessing Normality —40 to less than 60

- Normality Tests: All four of the tests for normality show p-values greater than $\alpha = 0.10$. Thus, we have evidence to suggest the x distribution is normal.
- QQ Plot: The data follows the agreement line with little deviation. This supports that the x distribution is normal.

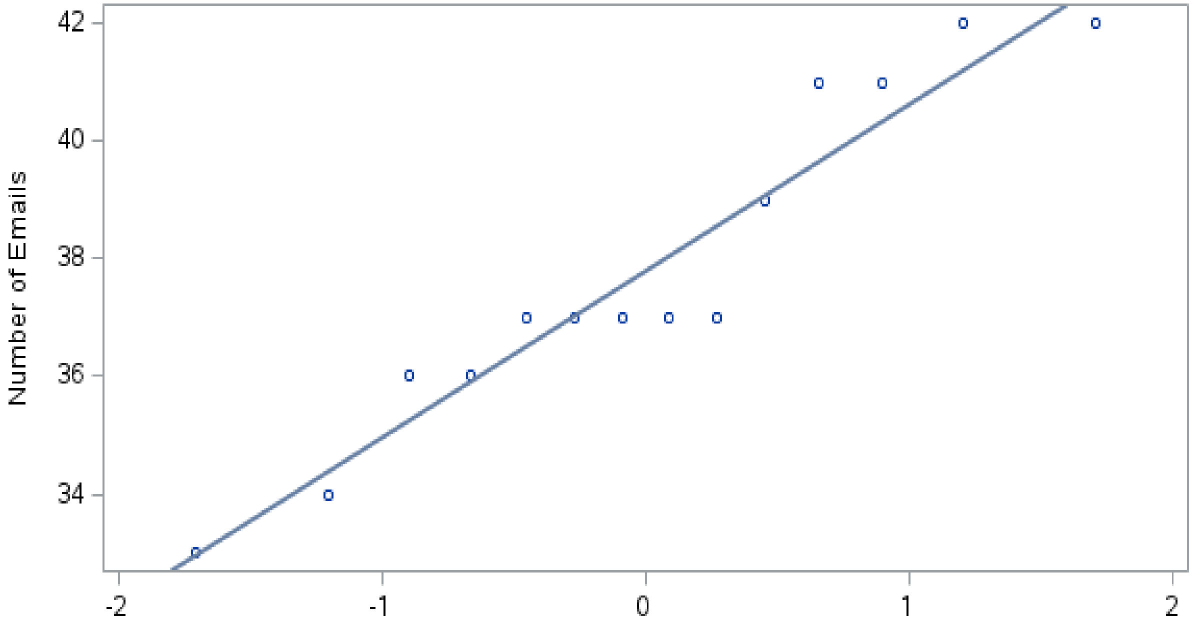
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.9345	Pr < W	0.2870
Kolmogorov-Smirnov	D	0.132574	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.049464	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.3612	Pr > A-Sq	>0.2500



Assessing Normality —60 and older

- **Normality Tests:** Three of the tests for normality show p-values less than $\alpha = 0.10$. Thus, we have evidence to suggest the x distribution is not normal.
- **QQ Plot:** The data follows the agreement line with a bit of deviation.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.912635	Pr < W	0.1721
Kolmogorov-Smirnov	D	0.252088	Pr > D	0.0169
Cramer-von Mises	W-Sq	0.111978	Pr > W-Sq	0.0728
Anderson-Darling	A-Sq	0.594986	Pr > A-Sq	0.0989



SAS Code: Assessing Homogeneity

```
/*A Rough Check for Homogeneity*/
TITLE "Table 1: Table to Compare Standard Deviations for Homogeneity";
PROC MEANS DATA = work.emails mean stddev VAR maxdec=4;
      class 'Age Group'n;
      VAR 'Number of Emails'n;
RUN;
TITLE;
```

Table 1: Table to Compare Standard Deviations for Homogeneity

The MEANS Procedure

Analysis Variable : Number of Emails				
Age Group	N Obs	Mean	Std Dev	Variance
Under 20	18	44.7778	2.9014	8.4183
20 to less than 40	14	41.2857	2.2678	5.1429
40 to less than 60	16	41.0625	2.7681	7.6625
60 and Older	14	37.7857	2.8333	8.0275

Assessing Homogeneity

- H_0: $\sigma^2_{\text{Under 20}} = \sigma^2_{\text{20 to less than 40}} = \sigma^2_{\text{40 to less than 60}} = \sigma^2_{\text{60 and older}}$

  H_A: At least one variance is different than the rest
- The ratio is less than 2. Thus, the standard deviations are close enough to use a test that requires homogeneity.

Ratio of standard deviations:

$$= \frac{SD_{under\ 20}}{SD_{20\ to\ less\ than\ 40}}$$

$$= \frac{2.9014}{2.2678}$$

$$= 1.2794 < 2$$

Choosing Hypothesis Test: Kruskal Wallis

- Since the data is not normal and homogeneous, we will perform the Kruskal Wallis test.
- The null hypothesis is that all age groups have the same median number of emails sent each month for the populations of all age groups.

H_0: $\eta_{Under\ 20} = \eta_{20\ to\ less\ than\ 40} = \eta_{40\ to\ less\ than\ 60} = \eta_{60\ and\ Older}$

- The alternative hypothesis is,

H_A: At least one age group has a different median number of emails sent in a month for the population.

- The level of significance, $\alpha = 0.10$, tells us that 10% of the time the analysis will conclude that at least one median is different when all medians are equal.

SAS Code: Kruskal Wallis and Post-hoc

```
/*Kruskal Wallis Test with Post Hoc DSCF*/
proc npar1way data=work.emails wilcoxon dscf;
    class 'Age Group'n;
    var 'Number of Emails'n;
run;
```

Wilcoxon Scores (Rank Sums) for Variable Number of Emails Classified by Variable Age Group

Age Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Under 20	18	856.50	567.0	64.186345	47.583333
20 to less than 40	14	429.00	441.0	59.124120	30.642857
40 to less than 60	16	473.50	504.0	61.875534	29.593750
60 and Older	14	194.00	441.0	59.124120	13.857143

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
28.1609	3	<.0001

The NPAR1WAY Procedure

Pairwise Two-Sided Multiple Comparison Analysis

Dwass, Steel, Critchlow-Fligner Method

Variable: Number of Emails

Age Group	Wilcoxon Z	DSCF Value	Pr > DSCF
Under 20 vs. 20 to less than 40	3.1695	4.4824	0.0083
Under 20 vs. 40 to less than 60	3.2269	4.5635	0.0069
Under 20 vs. 60 and Older	4.3298	6.1233	<.0001
20 to less than 40 vs. 40 to less than 60	0.2522	0.3566	0.9944
20 to less than 40 vs. 60 and Older	3.0241	4.2767	0.0133
40 to less than 60 vs. 60 and Older	2.8674	4.0552	0.0215

Performing Kruskal Wallis Test

- χ^2: 28.16 measures the variation between the average rank for each group and the overall average rank.
- P-value: There is a less than 0.01% chance of getting these ranks for the four age groups when all age groups have the same median number of emails sent in a month.
- Conclusion: Since less than 0.01% is less than 10%, we reject H_0. We are 90% confident that at least one age group has a different median of number of emails sent in a month.

Wilcoxon Scores (Rank Sums) for Variable Number of Emails Classified by Variable Age Group

Age Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Under 20	18	856.50	567.0	64.186345	47.583333
20 to less than 40	14	429.00	441.0	59.124120	30.642857
40 to less than 60	16	473.50	504.0	61.875534	29.593750
60 and Older	14	194.00	441.0	59.124120	13.857143

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
28.1609	3	<.0001

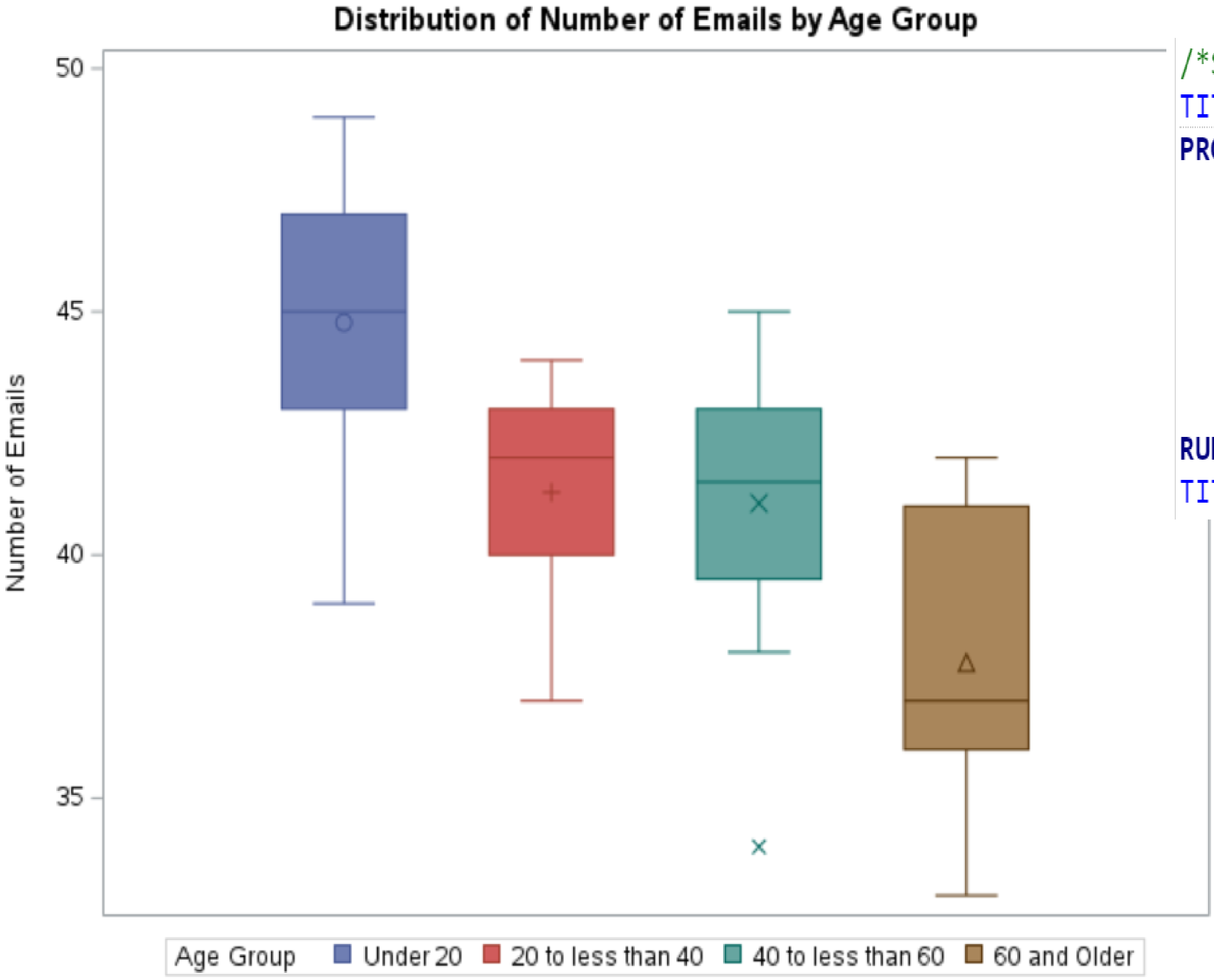
Post-hoc Tests: Dwass-Steel-Critchlow-Fligner Test

- Conclusion: The age group under 20 and the age group 60 and older have the smallest p-value and caused the significant Kruskal Wallis test at $\alpha = 0.10$.

The NPAR1WAY Procedure

Pairwise Two-Sided Multiple Comparison Analysis			
Dwass, Steel, Critchlow-Fligner Method			
Variable: Number of Emails			
Age Group	Wilcoxon Z	DSCF Value	Pr > DSCF
Under 20 vs. 20 to less than 40	3.1695	4.4824	0.0083
Under 20 vs. 40 to less than 60	3.2269	4.5635	0.0069
Under 20 vs. 60 and Older	4.3298	6.1233	$\boxed{<.0001}$
20 to less than 40 vs. 40 to less than 60	0.2522	0.3566	0.9944
20 to less than 40 vs. 60 and Older	3.0241	4.2767	0.0133
40 to less than 60 vs. 60 and Older	2.8674	4.0552	0.0215

Supporting Graphic: Stratified Box Plot




```
/*Stratified Box Plot*/
TITLE 'Box Plot for Number of Emails Sent by Age Group';
PROC sgplot data= work.emails;
    vbox 'Number of Emails'n / group= 'Age Group'n;
    title 'Distribution of Number of Emails by Age Group';
    yaxis label= 'Number of Emails';
    xaxis label= 'Age Group';
    ODS graphics
        /   attrpriority=none;
RUN;
TITLE;
```

Taking Action

- Further investigation is recommended.
- People under 20 may have more of the positions that require more email communication. It could also be the case that they may not be sending efficient or necessary emails. Company may need to provide lessons to less-experienced younger employees.
- People over 60 may have positions that require less email communication. It could also be the case that they send more efficient emails or prefer phone calls. Company may need to encourage older employees to use email as the world is becoming more virtual.

SAS Code: Screen Recording