Miza Syafiqah and Phichchaya Sutaporn

November 1st, 2022

1. Project Description

Sesame Street is a famous educational TV show in the United States. The company is interested in finding whether watching sesame street improves children's knowledge of numbers and letters. The data is collected by a third-party company and the company collected the data randomly across the United States. Biasness is minimized by making sure all the children involved are not in kindergarten and there are no children with special needs. The children took tests before and after watching sesame street for 6 months. The data collection portion of the study has been completed by a third-party company and the manager of Sesame Street has requested that STAT470W students to assists her in formally answering the research questions using the data.

1.1. Research Question

Question 1: Are there any differences in the view categories proportions of children that watch sesame street across the different sites?

Question 2: Is watching sesame street associated with children's knowledge of number improvement, on average? If so, how often do they have to watch?

Question 3: Is watching sesame street associated with children's knowledge of letter improvement, on average? If so, how often do they have to watch?

1.2 Variables

Pre-test of numbers and letters results are collected before the children watch sesame street. After 6 months of watching sesame street, post-test of numbers and letter results are collected. Children's living conditions (site) are also recorded to see whether it is associated with how many times the children watch sesame street. The sex and age of the children are also collected as well as where they watch sesame street (at home or preschool). Measurement of how well children understand the English language is also collected during the pre-test. A summary of all variables considered is included in Table 1

2. Exploratory Data Analysis (EDA)

It is always a good idea to define the variables used in the analysis before doing the analysis and answering the research question.

Variable	Type	Description	Level or range
site	Categorical	Children's living conditions	1: disadvantaged children who live in inner cities 2: advantaged suburban children 3: advantaged rural children 4: disadvantaged rural children
sex_coded	Categorical	Children's biological sex	1: male and 2: female

age	Numerical	Children's age (reported in months)	34 to 69 months
viewcat	Categorical	How often children watch sesame street in one week	1: rarely, 2: 1-2 times, 3: 3-5 times, and 4: more than 5 times
setting	Categorical	Where do children watch sesame street?	Preschool and Home
prenumb	Numerical	Pretest score for number test (out of 100)	1 to 52
prelet	Numerical	Pretest score for letter test (out of 100)	1 to 48
postnumb	Numerical	Posttest score for number test (out of 100)	0 to 54
postlet	Numerical	Posttest score for letter test (out of 100)	0 to 63
pptv	Numerical	Measurement of how well children understand English language (out of 100)	8 to 99
numbscore	Numerical	Measurement of knowledge of number improvement (postnumb-prenumb)	-13 to 33
letterscore	Numerical	Measurement of knowledge of letter improvement (postlet-prelet)	-17 to 41

Table 1: Summary of variables used in the analysis

Since both of numbscore and letterscore are our response variables for the second and third research question. First, let's take a closer look at the numbscore variable. Table 2 shows that there is a total of 222 observation units with no missing value. The minimum of numbscore is -35 while the maximum is 33. The mean of numbscore is 9.077 with a standard deviation of 9.077 while the median is 9 since the value of the mean is close to the median this suggests that the data is normally distributed.

Variable	N	N*	Mean	StDev	Minimum	Median	Maximum
numbscore	222	0	9.077	9.791	-35.000	9.000	33.000

Table 2: Summary statistics for the numbscore variable.

Second, looking at summary statistics for the letterscore variable from table 3 suggests that there is a total of 222 observation units with no missing value. The minimum of letterscore is -22 while the maximum is 41. The mean of letterscore is 10.896 with a standard deviation of 11.347 while the median is 9 since the mean value is a bit larger than the median this suggests that the data is slightly right-skewed.

Variable	N	N*	Mean	StDev	Minimum	Median	Maximum
letterscore	222	0	10.896	11.347	-22.000	9.000	41.000

Table 3: Summary statistics for the letterscore variable.

We created the EDAs for three research questions.

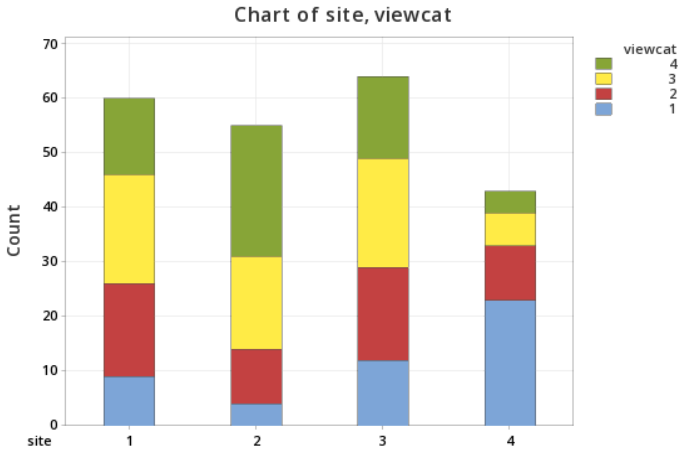

Figure 1: Stack boxplot of view category and site

From figure1, firstly in site1, we can see that viewcat3 has the highest proportion. Secondly, in site2, viewcat 4 has the highest proportion. Thirdly, in site3, viewcat3 has the highest proportion. Lastly, in site 4, viewcat1 has the highest proportion. This EDA suggests that for each site there might be differences in the proportions of children that watch sesame street accounting for each view category.

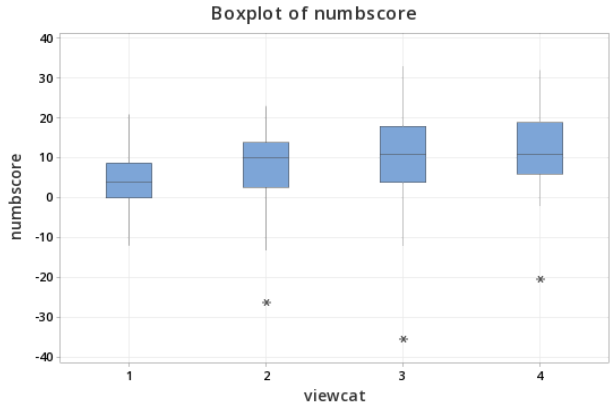

Figure 2: Box plot of the difference between prenumb and postnum (numbscore)

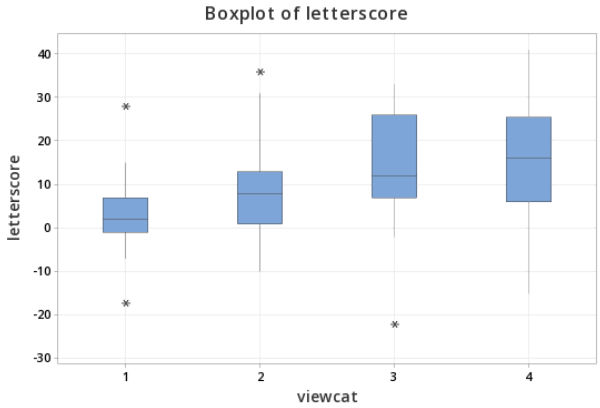

Figure 3: Box plot of the difference between prelet and postlet (letterscore)

From figures 2 & 3, the average numbscore and letterscore are all positive. We also created 2 more EDAs (figure 6 and 7 in Appendix) and they suggest that for every viewcat, postnumb and postlet have a higher average than prenumb and prelet. All EDAs suggest that there might exist improvements in children's knowledge of numbers and letters after watching Sesame Street. Note that the purpose of EDAs is only to give audiences some idea and a better understanding of data. However, only looking at EDAs cannot answer the research question yet. Therefore, to answer the research questions, further statistical analysis needs to be conducted.

3. Statistical Analysis

For the first question, to determine whether are there any differences in the view categories proportions of children that watch sesame street across the different sites. We performed the Chi-square test of homogeneity

	Chi-Square	DF	P-Value
Pearson	43.930	9	0.000
Likelihood Ratio	41.006	9	0.000

Table 4: Table of Chi-square test of homogeneity

Since the both p-values of Pearson and the Likelihood Ratio test are significant (less than 0.05), we reject the null hypothesis. There is sufficient evidence at the 0.05 level to conclude that proportion of children who watch sesame street differs among sites and view categories.

For the second question, to determine whether watching sesame street is associated with children's knowledge of numbers improvement, on average or not, and if so, how often they have to watch. We fitted the ANCOVA model with numbscore as a response; site, sex, viewcat, and setting as factors; age and ppvt as covariates and two-way interaction terms. We found that there are most three extreme outliers (see table 9 in the Appendix), so we decided to remove those three outliers because the assumptions of normality would not be met and the standardized residuals are beyond $\pm$ 2. After getting rid of them, per residual plots (figure 9 in the Appendix), the model assumptions are satisfied. Then we used the backward elimination with alpha equal to 0.05 and arrived at our final model below:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
viewcat	3	2275.7	758.56	10.99	0.000
Error	215	14840.0	69.02		
Lack-of-Fit	213	14597.5	68.53	0.57	0.827
Pure Error	2	242.5	121.25		
Total	218	17115.7			

Table 5: ANCOVA table for the final model for the second research question

As we can see in table 5, viewcat is the only variable that is significant (p-value is less than 0.05) meaning that viewcat does associate with children's knowledge of numbers improvement. Furthermore, we answered the second part of the question about how often children have to watch sesame street by performing the Tukey comparison and got the result below:

Grouping Information Using the Tukey Method and 95% Confidence

viewcat	N	Mean	Grouping
4	56	12.8036	A
3	62	11.5484	A
2	53	8.7547	A
1	48	4.1458	B

Table 6: Tukey Method of comparison result

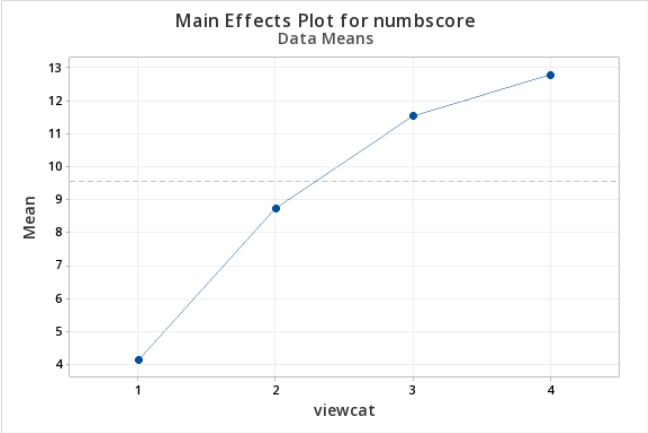

Figure 4 : Main Effects Plot for numbscore

From table 6 and figure 4, we can see that average of numbscore for viewcat 4,3 and 2 are significantly higher than viewcat1 meaning that watching sesame street at least 1 time per week associate with children's knowledge of numbers improvement.

Similarly, the third question, to determine whether watching sesame street is associated with children's knowledge of letters improvement, on average or not, and if so, how often they have to watch. We fitted the ANCOVA model with letterscore as a response; site, sex, viewcat, and setting as factors; age and ppvt as covariates and two-way interaction terms. We found that there are most two extreme outliers (see table 10 in the Appendix), so we decided to remove those two outliers because the assumptions of normality would not be met and the standardized residuals are beyond $\pm$ 2. The two outliers are similar to two of the previous model's outliers so it is worth to look what happened to the students when they are taking the tests. After getting rid of them, per residual plots (figure 11) in the Appendix, the model assumptions are satisfied. Then we used the backward elimination with alpha equal to 0.05 and arrived at our final model below:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
age	1	446.3	446.34	6.11	0.014
site	3	485.7	161.91	2.22	0.087
viewcat	3	4976.7	1658.89	22.70	0.000
age*site	3	626.8	208.94	2.86	0.038
Error	209	15275.5	73.09		
Lack-of-Fit	207	14947.5	72.21	0.44	0.894
Pure Error	2	328.0	164.00		
Total	219	26688.1			

Table7: ANCOVA table for the final model for the third research question

As we can see in table 7, viewcat variable is significant (p-value is less than 0.05) meaning that viewcat does associate with children's knowledge of letters improvement. We included age, site, and interaction between age and site because it is significant. They affect the relationship between letterscore and viewcat. Furthermore, we answered the second part of the question about how often children have to watch sesame street by performing the Tukey comparison and got the result below:

Grouping Information Using the Tukey Method and 95% Confidence

viewcat	N	Mean	Grouping		
3	62	16.1208	A		
4	56	15.7433	A		
2	54	8.5651		B	
1	48	2.9534			C

Table 8: Tukey Method of Comparison for the third research question

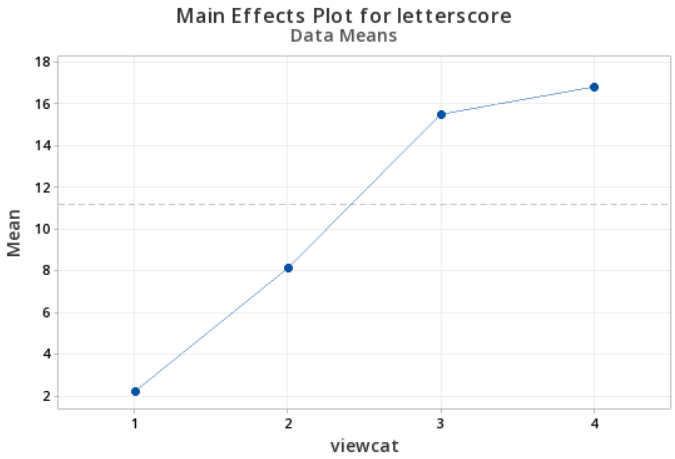

Figure 5: Main Effects Plot for letterscore

From table 8 and figure 5, we can see that average of letterscore for viewcat 4 and 3 are significantly higher than viewcat 1 and 2 meaning that watching sesame street at least 3 times per week associate with children's knowledge of numbers improvement.

4. Recommendations

The described analysis of your data results in the recommendations below:

Question 1: Are there any differences in the view categories proportions of children that watch sesame street across the different sites?

Yes, there are differences in the view categories proportions of children that watch sesame street across the different sites. Each site has a different viewcats proportion.

Question 2: Is watching sesame street associated with children's knowledge of number improvement, on average? If so, how often do they have to watch?

Yes, watching sesame street does associate with children's knowledge of number improvement, on average. viewcat 2, 3, and 4 have a higher average than viewcat 1. Thus, watching sesame street at least one time per week is associated with children's knowledge of numbers improvement.

Question 3: Is watching sesame street associated with children's knowledge of letter improvement, on average? If so, how often do they have to watch?

Yes, watching sesame street does associate with children's knowledge of letter improvement, on average. viewcat 3 and 4 have a higher average than viewcat 1 and 2. Thus, watching sesame street at least three times per week is associated with children's knowledge of letters improvement.

5. Resources

For resources related to the ANCOVA, see https://online.stat.psu.edu/stat502_fa21/lesson/9/9.3
For resources related to the Chi-Square Test, see https://online.stat.psu.edu/stat500/lesson/8

6. Additional Considerations

We answered your first research question using a Chi-Square test for Independence and both of your second and third research questions using ANCOVA and Tukey comparison methods. In all cases, the conditions of the model were reasonably met and the results should be considered trustworthy. However, as you might see in the Appendix part the R-squared is not so high. Considering more variables might help improve the R-squared. For example, parents' marital status may associate with how much children can improve their number and letter knowledge. Also, it is worth looking further at the outliers as the children maybe are not in the best state when they took the test. For supporting figures, please see the Appendix. The Tukey comparison plot that is located in the Appendix shows the significance between interaction variables and if you want us to elaborate, please do not hesitate to contact us if you have any questions. There is also a caution below:

Association is not causation: This is an observational study, so we cannot make cause-and-effect conclusions. For example, we can only say that watching sesame street is associated with children's knowledge of numbers improvement but we cannot conclude that watching sesame street causes children's knowledge of numbers improvement.

Technical Appendix

Further EDA

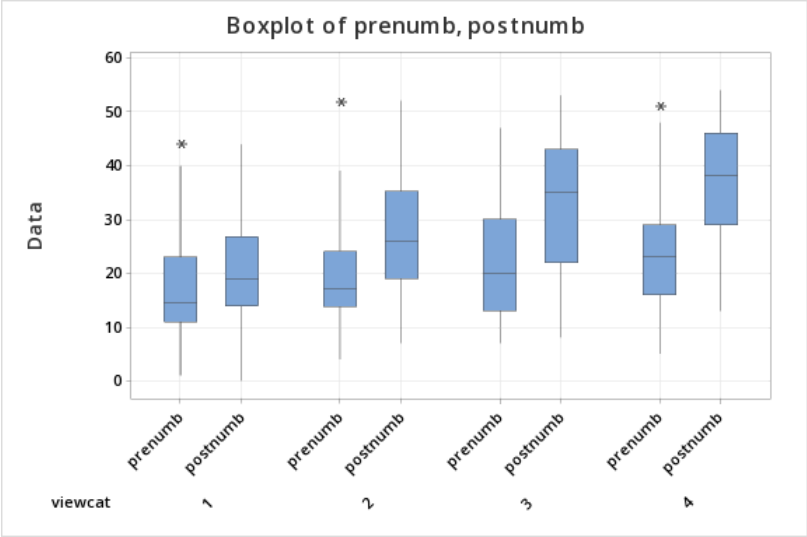

Figure 6: Boxplot of prenumb and postnumb

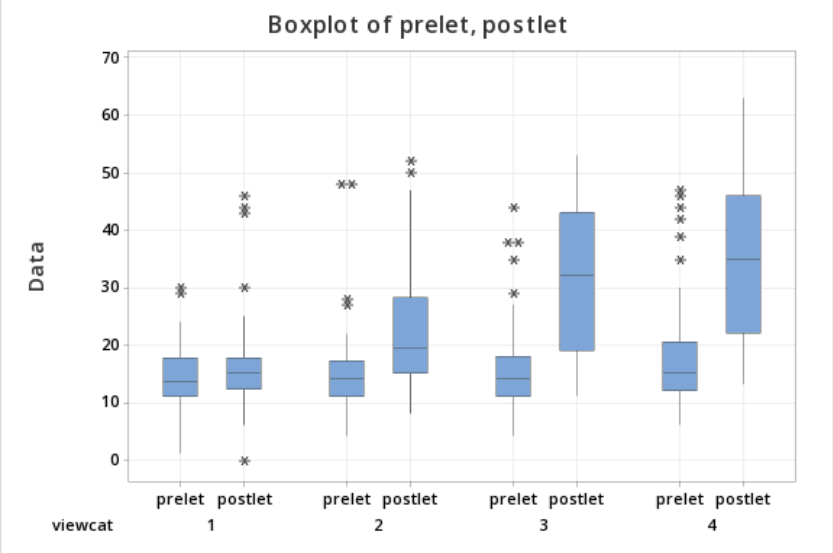

Figure 7: Boxplot of prelet and postlet

Id	Site	Sex	Age	Viewcat	Setting	Numbscore
27	1	male	61	4	preschool	-20
40	1	male	51	2	home	-26
50	1	male	49	3	home	-35

Table 9: Outliers table for the second research question

Id	Site	Sex	Age	Viewcat	Setting	Numbscore
27	1	male	61	4	preschool	-20
50	1	male	49	3	home	-35

Table 10: Outliers table for the third research question

Assessing ANOVA for the second research question:

1. **First full model before getting rid of outliers:**

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
ppvt	1	520.4	520.37	6.08	0.014
site	3	809.4	269.79	3.15	0.026
viewcat	3	1569.4	523.13	6.12	0.001
Error	214	18305.5	85.54		
Lack-of-Fit	212	18063.0	85.20	0.70	0.757
Pure Error	2	242.5	121.25		
Total	221	21187.7			

Table 11: ANCOVA table for the second research question before getting rid of outliers

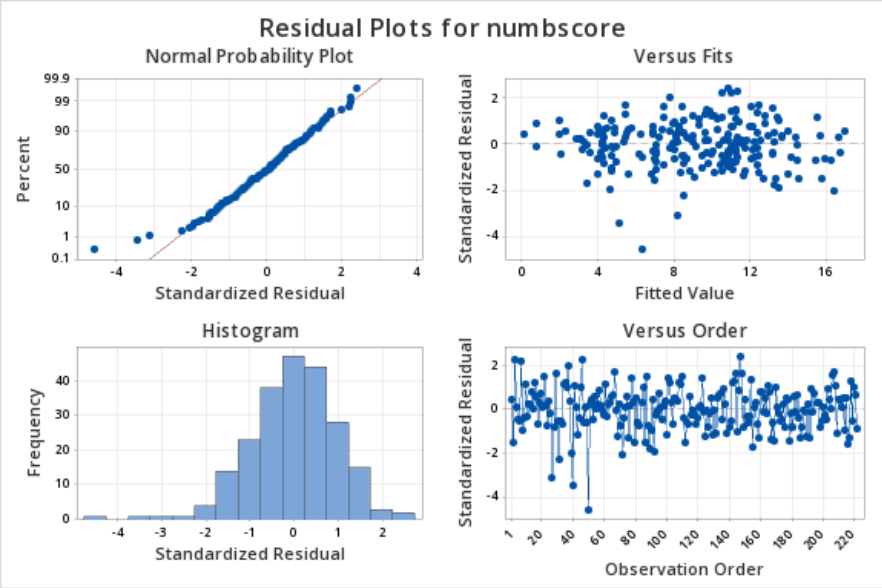

Figure 8: Residual plots for numbscore before removing three outliers

2. **The second and final model after getting rid of 3 outliers can be found in the Statistical Analysis section and the residual plots are as below:**

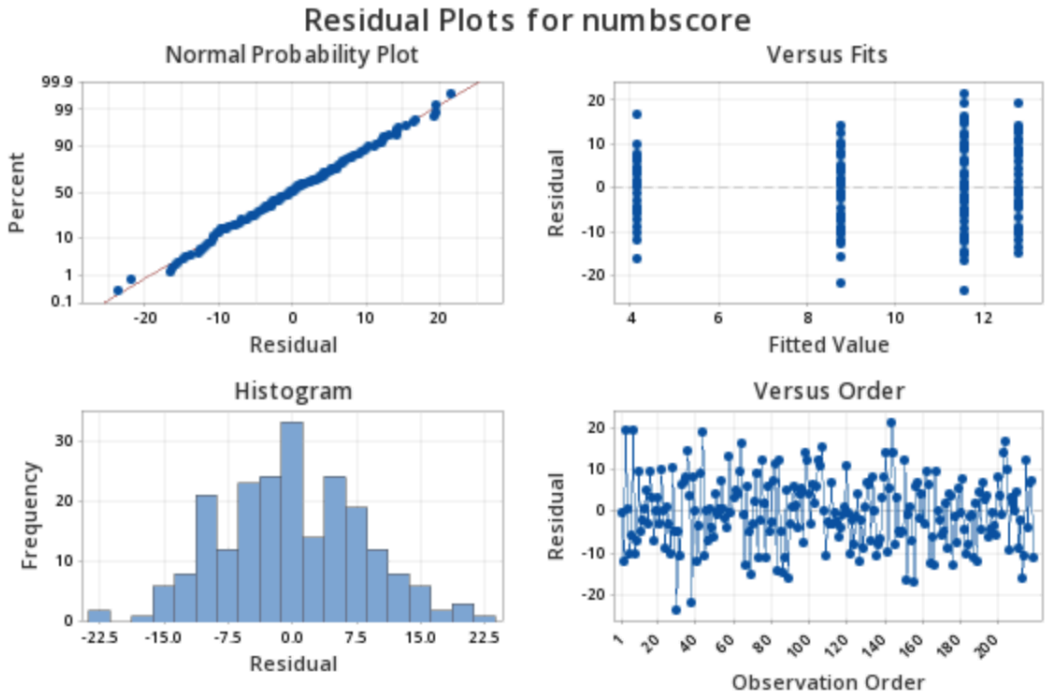

Figure 9: Residual plots fore numbscore after removing three outliers

Getting rid of 3 outliers makes the normal probability plot look better. Now, all points do follow the red line indicating that the normally distributed assumption is satisfied. The assumption of equal variance is met because there is no cone pattern in the residual vs fit plot. Therefore, we decided to remove these three outliers.

Assessing ANOVA for the third research question:

1. **First full model before getting rid of outliers:**

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
age	1	139.4	139.40	1.74	0.189
ppvt	1	321.0	321.04	4.01	0.047
sex	1	340.2	340.22	4.25	0.041
site	3	2227.3	742.44	9.27	0.000
viewcat	3	4271.0	1423.68	17.77	0.000
setting	1	8.0	8.02	0.10	0.752
age*ppvt	1	311.8	311.79	3.89	0.050
ppvt*sex	1	499.7	499.71	6.24	0.013
viewcat*setting	3	1059.4	353.12	4.41	0.005
Error	206	16502.8	80.11		
Lack-of-Fit	204	16174.8	79.29	0.48	0.871
Pure Error	2	328.0	164.00		
Total	221	28456.6			

Table 12: ANCOVA table for the third research question before getting rid of outliers

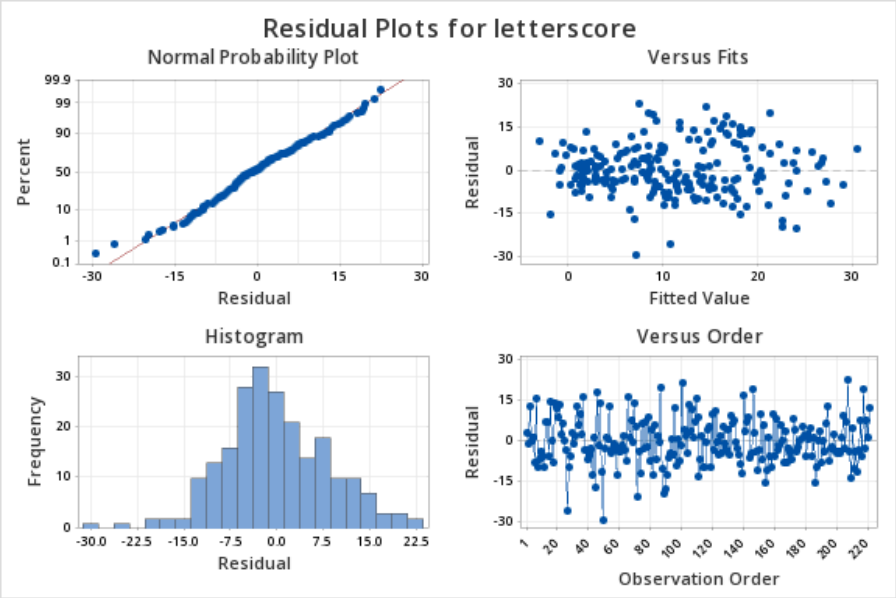

Figure 10: Residual plots for letterscore before removing three outliers

2. **The second and final model after getting rid of 2 outliers can be found in the Statistical Analysis section and the residual plots are as below:**

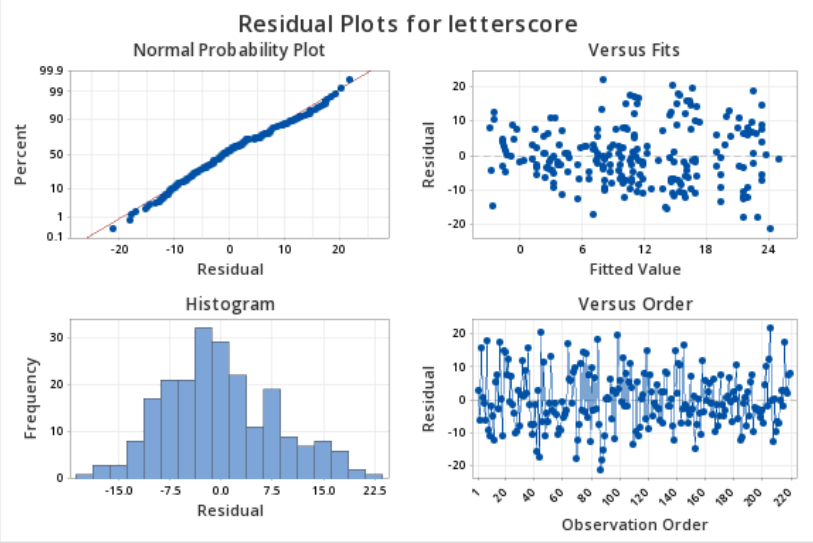

Figure 11: Residual plots for letterscore before removing three outliers

Getting rid of 2 outliers makes the normal probability plot look better. Now, all points do follow the red line indicating that the normally distributed assumption is satisfied. The assumption of equal variance is met because there is no cone pattern in the residual vs fit plot. Therefore, we decided to remove these two outliers.