Problem 1

A national chain of women's clothing stores with locations in large shopping malls thinks that it can do a better job of planning more renovations and expansions if it understands what variables impact sales. It plans a small pilot study on stores in 25 different mall locations. The data it collects consists of monthly sales, store size (sq. ft), number of linear feet of window display, number of competitors located in the mall, size of the mall (sq. ft), and distance to the nearest competitor (ft).

In this report, we will analyze sales data from a national chain of women's clothing stores and determine what variables impact sales. The data has six variables described below and 25 records.

Variable	Description
Sales	Number of Sales per store
Size	Store size in square feet
Windows	Number of linear feet of window display
Competitors	Number of competitors in the mall
Mall Size	Mall size in square feet
Nearest Competitor	Distance between the store and the nearest competitor

a. Find a multiple regression model for the data.

Sales is the dependent variable and the rest are independent variables that have an impact on sales with the assumption of **linearity, Normality, Independence, and Homoscedasticity** of the data for all variables we write our initial regression line as the following

Sales = β_0 + β_1 Size + β_2 Windows + β_3 Competitors + β_4 Mall Size + β_5 Nearest Competitor

b. Interpret the values of the coefficients in the model.

By running the SAS Regression procedure we get the following values for the coefficients of the variables

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1506.80179	672.18680	2.24	0.0371
Size	Size	1	0.91937	0.30063	3.06	0.0065
Windows	Windows	1	9.07598	28.82343	0.31	0.7563
Competitors	Competitors	1	-67.68553	21.95288	-3.08	0.0061
Mall_Size	Mall_Size	1	-0.00090285	0.00028062	-3.22	0.0045
Nearest_Competitor	Nearest_Competitor	1	2.09589	1.59443	1.31	0.2043

From the previous table

- The Intercept (β_0) is equal to 1506.8
- The Regression Coefficient of Size (β_1) is equal to 0.91937
- The Regression Coefficient of Windows (β_2) is equal to 9.07598
- The Regression Coefficient of Competitors (β_3) is equal to -67.68553
- The Regression Coefficient of Mall_Size (β_4) is equal to -0.0009
- The Regression Coefficient of Nearest_Competitior (β_5) is equal to 2.09589

Thus our Regression line is interpreted as the following

Sales = 1506.8 + 0.91937 Size + 9.07598 Windows - 67.68553 Competitors
- 0.0009 Mall Size + 2.096 Nearest Competitor

Both Competitors and Mall_size have a negative effect on Sales even though Mall_size's effect is neglectable.

c. Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion?

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5761406	1152281	19.21	<.0001
Error	19	1139390	59968		
Corrected Total	24	6900796			

The Regression analysis suggests that our model is significant and the P value is <.0001 which is less than α (0.05).

d. Use the model to predict monthly sales for each of the stores in the study.

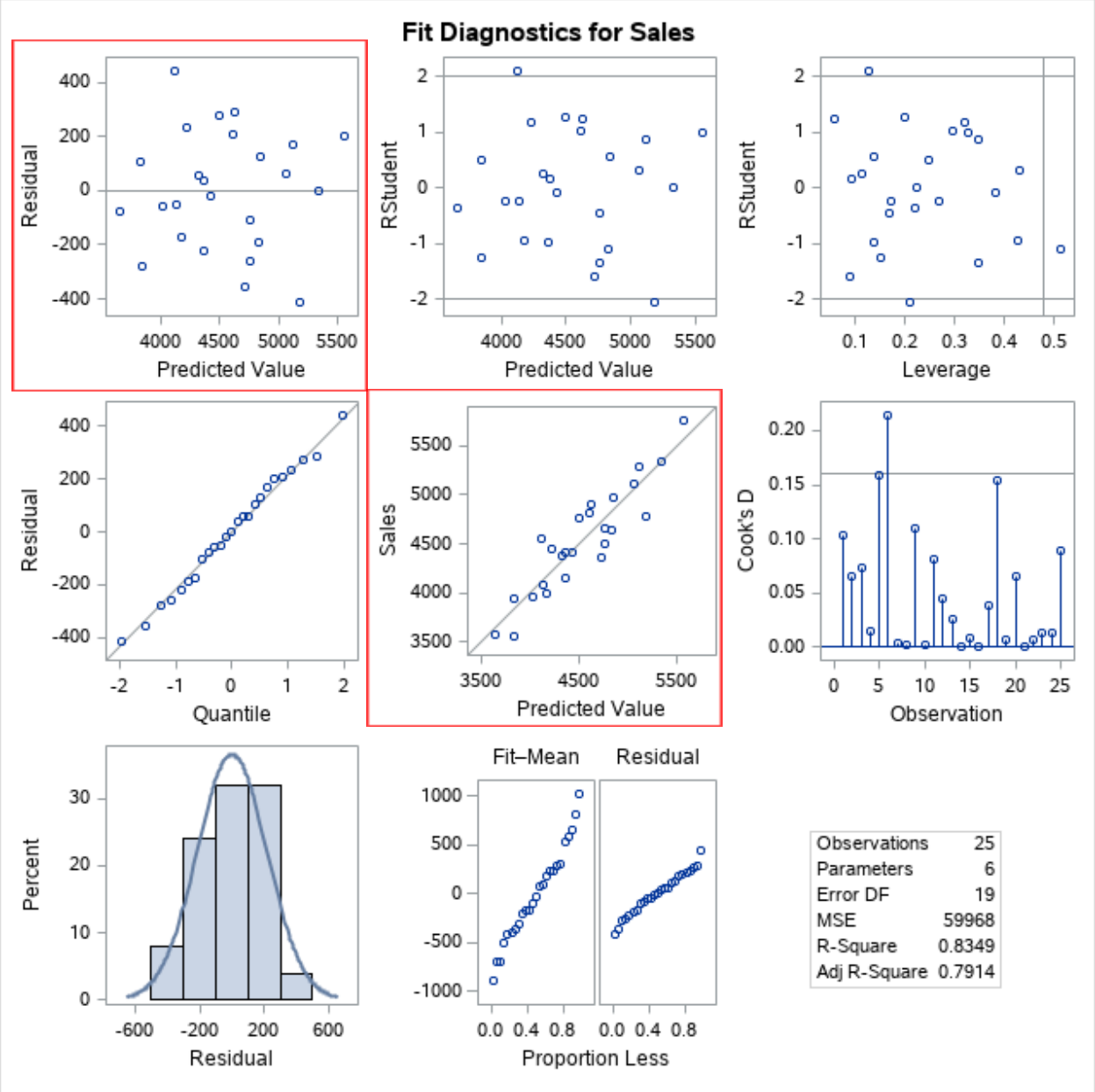
By looking at the results of the Regression procedure, the table below shows the predicted values based on the previous model, and from first glance, the predicted values are close to the actual sales.

Output Statistics							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	4453	4221	138.4843	231.9397	202.0	1.148	0.103
2	4770	4496	109.7095	274.1233	218.9	1.252	0.066
3	4821	4611	133.3674	209.6974	205.4	1.021	0.073
4	4912	4625	58.2133	287.1634	237.9	1.207	0.015
5	4774	5188	111.9124	-413.8542	217.8	-1.900	0.159
6	4638	4827	175.4666	-188.5227	170.8	-1.104	0.214
7	4076	4129	126.7285	-52.6221	209.5	-0.251	0.004
8	3967	4025	101.6295	-57.7434	222.8	-0.259	0.002
9	4000	4175	160.0566	-174.6896	185.3	-0.943	0.110
10	4379	4321	82.4775	57.7525	230.6	0.250	0.001
11	5761	5559	139.8824	201.5490	201.0	1.003	0.081
12	3561	3839	94.8106	-277.8760	225.8	-1.231	0.045
13	4145	4366	90.8745	-221.3260	227.4	-0.973	0.025
14	4406	4369	74.3130	36.9650	233.3	0.158	0.000
15	4972	4844	90.5548	128.3427	227.5	0.564	0.008
16	4414	4433	151.4788	-18.8276	192.4	-0.098	0.001
17	4363	4721	72.7112	-358.2757	233.8	-1.532	0.038
18	4499	4759	144.7183	-259.6539	197.5	-1.314	0.155
19	3573	3651	114.4841	-78.2762	216.5	-0.362	0.006
20	5287	5116	144.0371	171.0334	198.0	0.864	0.066
21	5339	5338	116.0869	1.4031	215.6	0.007	0.000
22	4656	4762	100.1693	-105.7615	223.5	-0.473	0.008
23	3943	3837	121.4905	106.1836	212.6	0.499	0.014
24	5121	5061	160.8775	59.6326	184.6	0.323	0.013
25	4557	4115	87.0493	441.6432	228.9	1.930	0.090

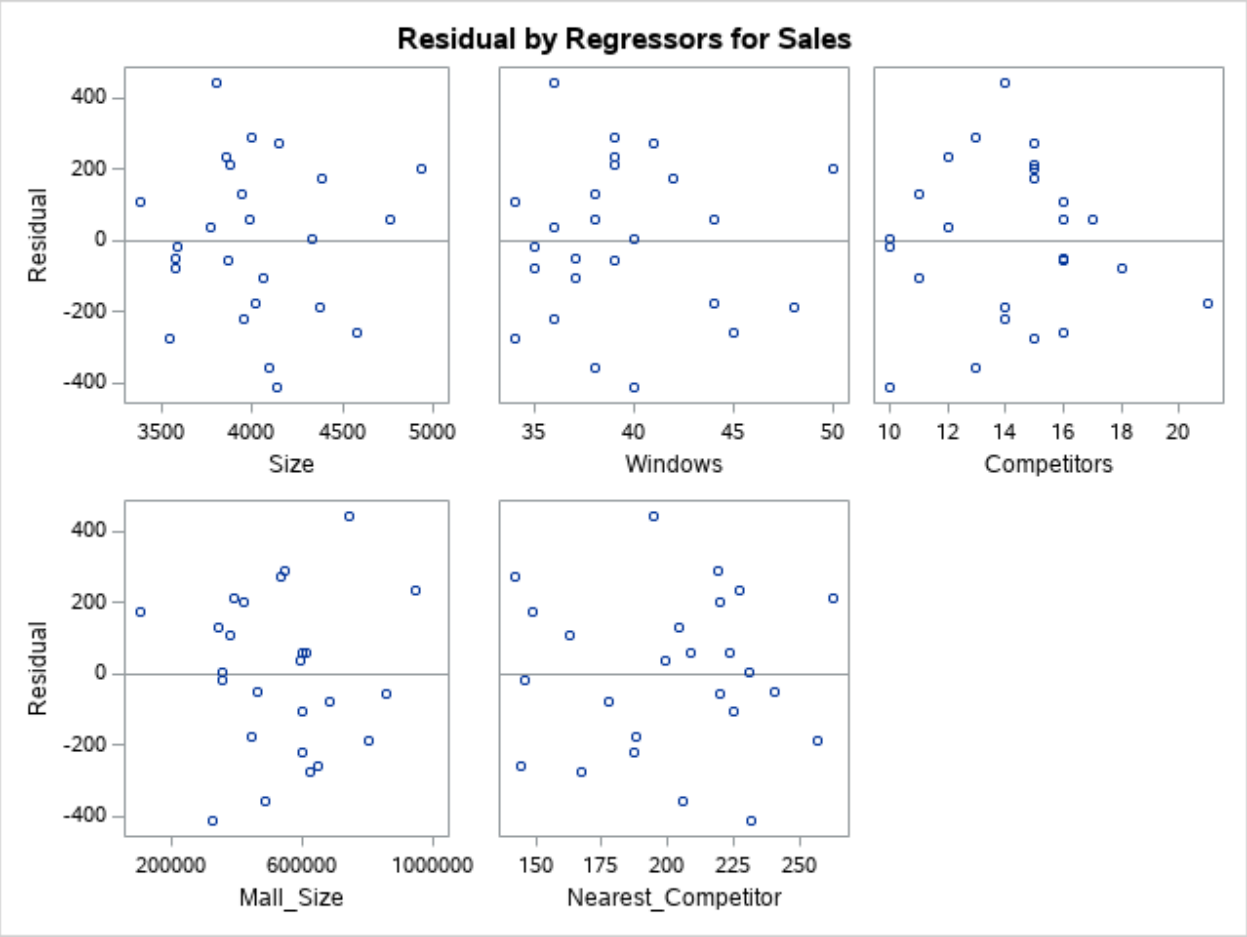
e. Plot the residuals versus the actual values. Do you think that the model does a good job of predicting monthly sales? Why or why not?

Yes, The model is doing a good job in predicting sales values because when Sales vs Predicted sales are plotted it is clear that the values are near the diagonal straight line. Also by looking at the residual vs predicted value plot all points are near the zero residual line.

It is worth mentioning that there is only 1 influential point (observation 6) with a cook score of 0.241



And there are no systematic patterns to and of the variables residual plots



f. Find and interpret the value of R^2! for this model.

Root MSE	244.88345	R-Square	0.8349
Dependent Mean	4535.48000	Adj R-Sq	0.7914
Coeff Var	5.39928		

R-Square is equal to 0.8349 which is accepted based on general industry norms.

g. Do you think that this model will be useful in helping planners? Why or why not?

Given the R-Square value of 0.8349, the P value of the model <.0001 that is less than α (0.05), Sales Vs predicted plot, and Residual by Regressors for Sales plots I would fairly assume that the model is useful in helping the planners with renovations and expansion.

h. Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions?

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1506.80179	672.18680	2.24	0.0371
Size	Size	1	0.91937	0.30063	3.06	0.0065
Windows	Windows	1	9.07598	28.82343	0.31	0.7563
Competitors	Competitors	1	-67.68553	21.95288	-3.08	0.0061
Mall_Size	Mall_Size	1	-0.00090285	0.00028062	-3.22	0.0045
Nearest_Competitor	Nearest_Competitor	1	2.09589	1.59443	1.31	0.2043

The individual regression coefficients suggest that variables Windows and Nearest_competitior are not statistically significant in their correlations with Sales, their P values of 0.7563 and 0.2043 respectively are both higher than α (0.05).

i. If you were going to drop just one variable from the model, which one would you choose? Why?

I would choose to drop the variable Windows because it has the highest P value for being not statistically significant that it correlates with Sales.

The store planners for the women's clothing chain want to find the best model that they can for understanding what store characteristics impact monthly sales.

j. Use stepwise regression to find the best model for the data.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1769.60574	611.03962	508470	8.39	0.0086
Size	1.04482	0.13276	3755185	61.94	<.0001
Competitors	-71.03060	18.90237	856069	14.12	0.0012
Mall_Size	-0.00079216	0.00027187	514713	8.49	0.0083

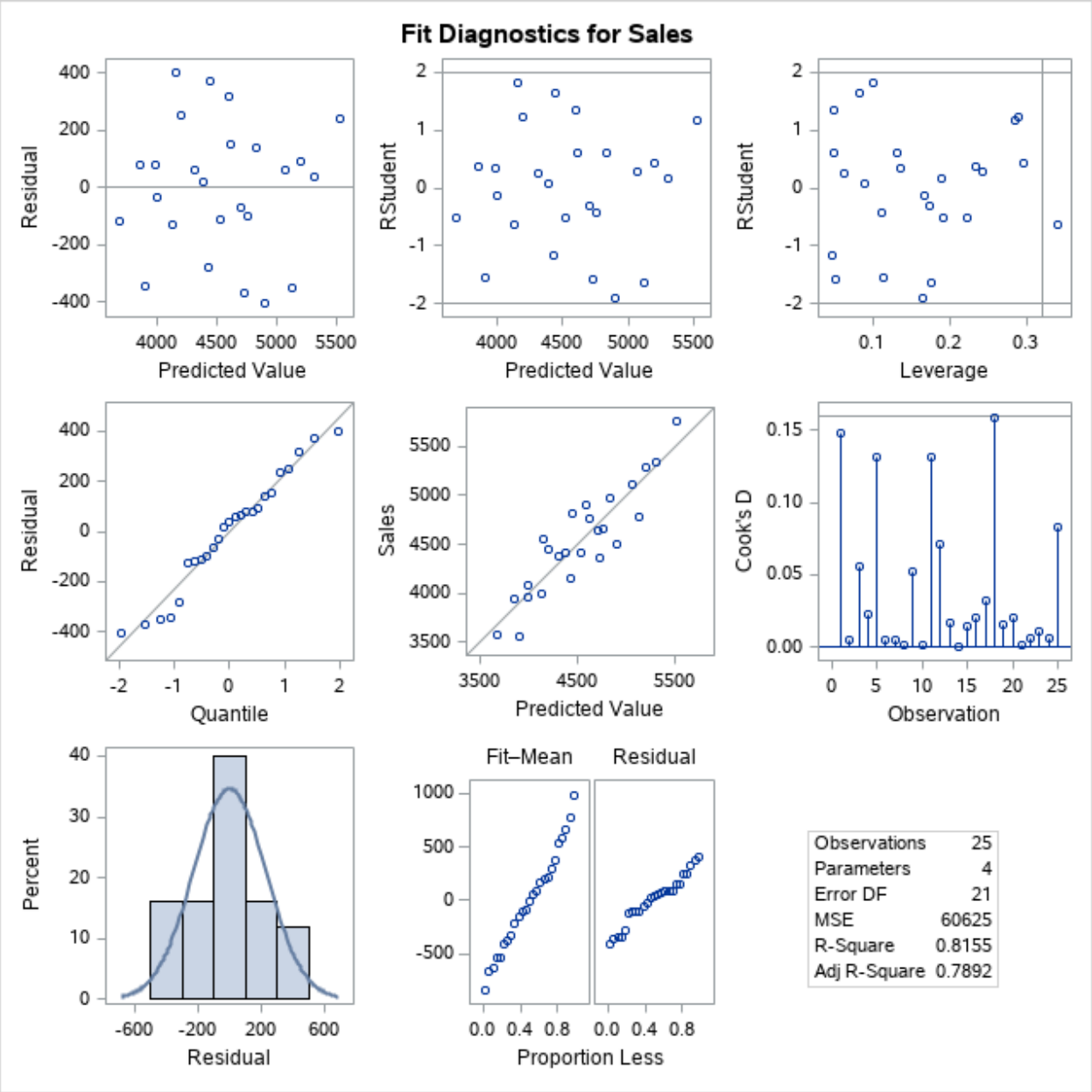
After 3 steps using the stepwise regression we are left with three variables Size (1.04482) Competitors (-71.0306) Mall_Size (-0.000792)

the model will look like the following

Sales = 1769.60574 + 1.04482 Size - 71.0306 Competitors - 0.000792 Mall Size

k. Analyze the model you have identified to determine whether it has any problems.

The new model has an R-square value of 0.8155 and all the variables are significant with a P value below α (0.05). There are no influential points and the points are close to the diagonal straight line when Sales vs predicted values are plotted, and there are no systematic patterns to and of the variables' residual plots.



I. Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen.

In this analysis, we tested two regression models and concluded that the following model is ideal for predicting sales based on three variables Size, Competitors, and Mall_size.

Sales = 1769.60574 + 1.04482 Size - 71.0306 Competitors - 0.000792 Mall Size

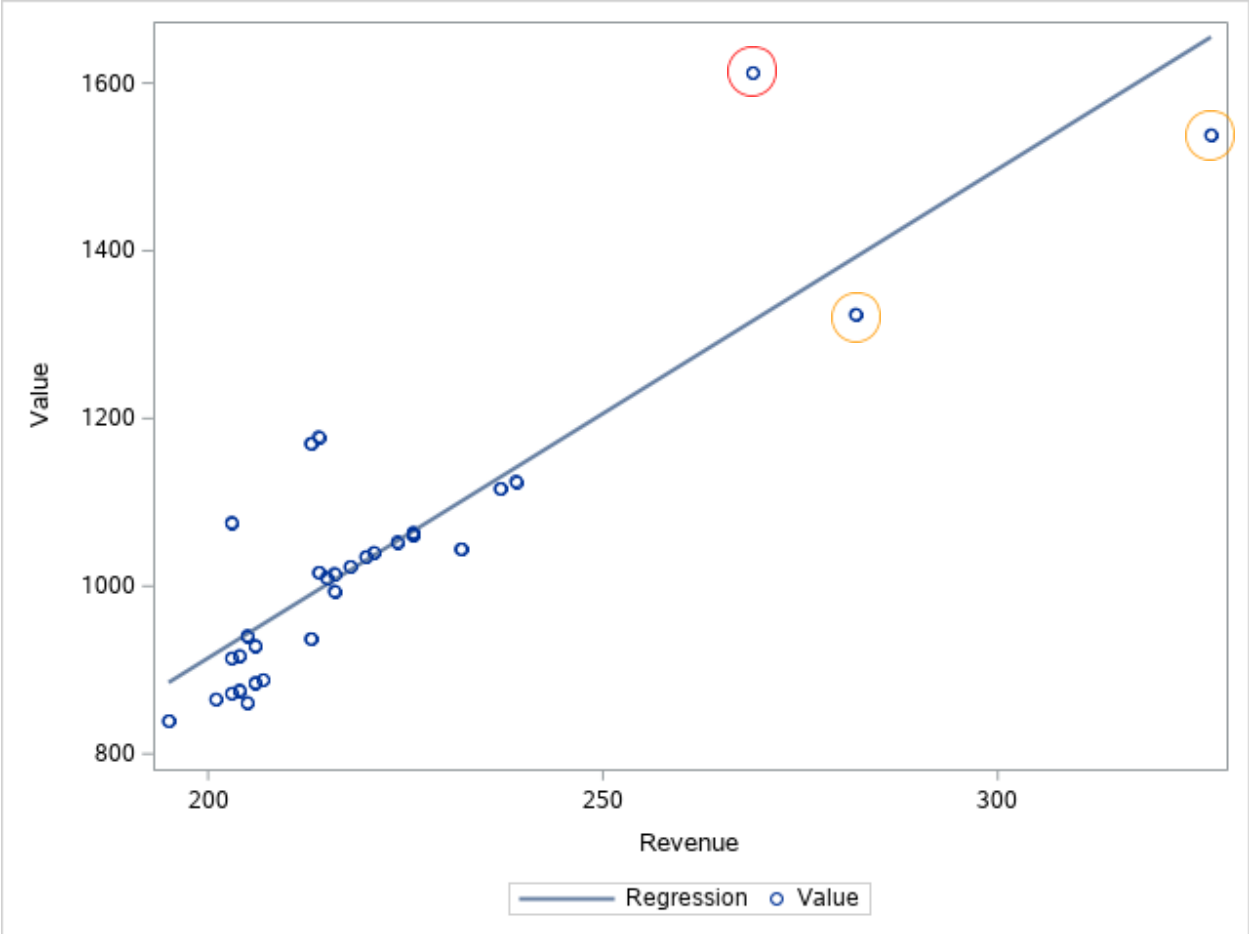
All variables are significant at α (0.05) and the R-square value is 0.8155 which is accepted by the industry norm and there are no influential points. In comparison, the other model had a higher R-square value of 0.8349 but two variables had a P value above α (0.05) and had an influential point.

Problem 2

The File NFLValues.xlsx shows the annual revenue ($ million) and the estimated team value ($ million) for the 32 teams in the National Football League.

a. Develop a scatter diagram with Revenue on the horizontal axis and Value on the vertical axis. Does it appear that there are any outliers and/or influential observations in the data?

According to the scatter plot, there are three outliers that are far from the majority of the data points. Two of them highlighted in orange are close to the regression line and one is far highlighted in red



b. Develop the estimated regression equation that can be used to predict team value given the value of annual revenue.

For the initial regression model, Value is the dependent variable and Revenue is the independent variable therefore our model is as follows

Value = β_0 + β_1 Revenue

After running the Regression analysis the intercept value is -252.0783 and the coefficient of value is 5.83167 and both the intercept and Revenue are significant at α (0.05), The coefficient of determination R^2 is equal to 0.7673 and it is not ideal according to the industry norm, most likely this is due to the presence of influential points.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-252.07830	130.81712	-1.93	0.0635
Revenue	Revenue	1	5.83167	0.58631	9.95	<.0001

Root MSE	87.24412	R-Square	0.7673
Dependent Mean	1040.00000	Adj R-Sq	0.7596
Coeff Var	8.38886		

From our findings, the Regressions model is as follows

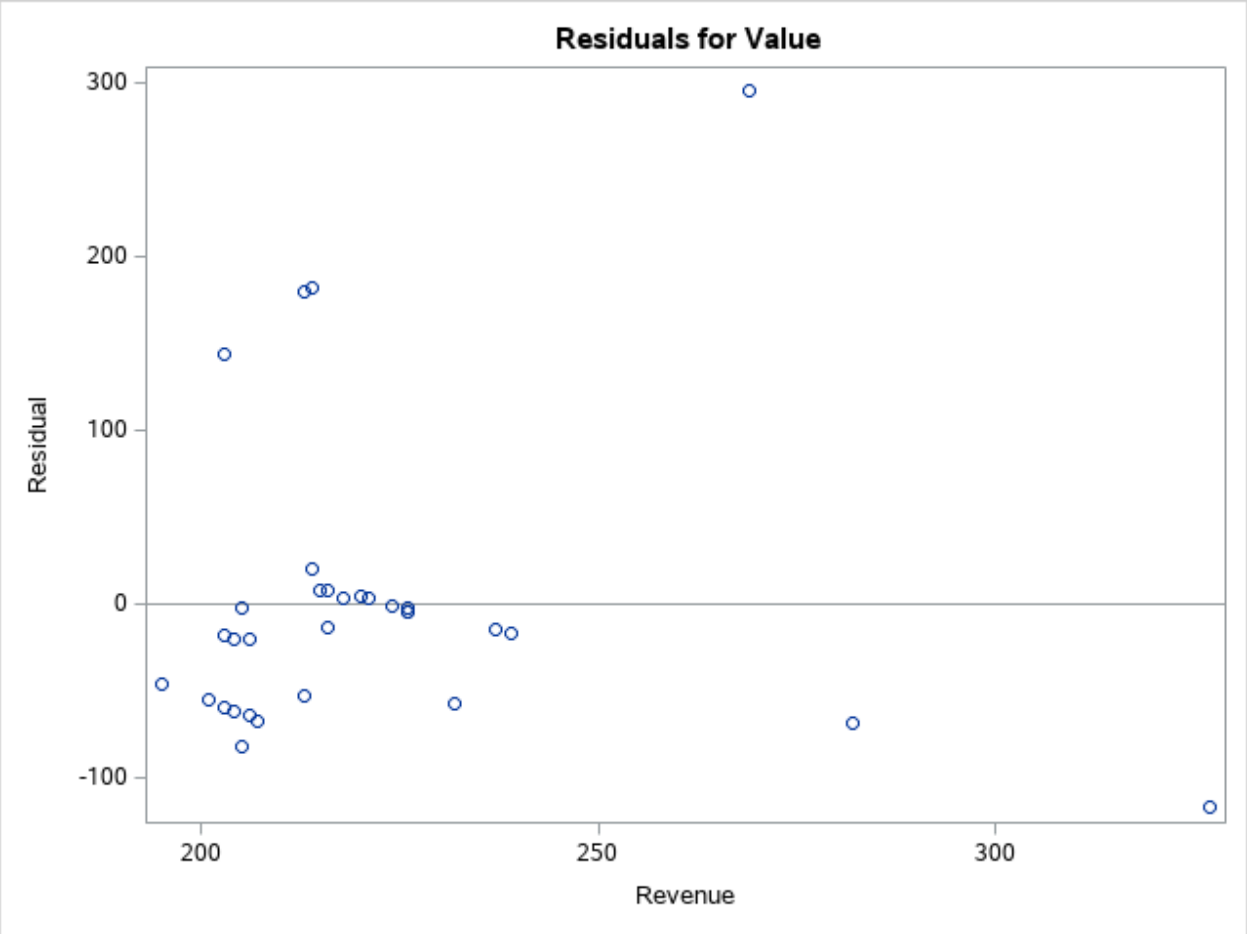
Value = -252.0783 + 5.83167 Revenue

from the equation we can predict the Values and test them against actual data points.

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	914	931.7497	18.8762	-17.7497	85.178	-0.208	0.001
2	872	931.7497	18.8762	-59.7497	85.178	-0.701	0.012
3	1062	1066	15.6406	-3.8780	85.831	-0.045	0.000
4	885	949.2447	17.9197	-64.2447	85.384	-0.752	0.012
5	1040	1037	15.4283	3.2803	85.869	0.038	0.000
6	1064	1066	15.6406	-1.8780	85.831	-0.022	0.000
7	941	943.4130	18.2253	-2.4130	85.319	-0.028	0.000
8	1035	1031	15.4499	4.1120	85.865	0.048	0.000
9	1612	1317	31.8030	295.3804	81.241	3.636	1.013
10	1061	1066	15.6406	-4.8780	85.831	-0.057	0.000
11	917	937.5814	18.5443	-20.5814	85.250	-0.241	0.001
12	1023	1019	15.5635	3.7753	85.845	0.044	0.000
13	1125	1142	18.5037	-16.6897	85.259	-0.196	0.001
14	1076	931.7497	18.8762	144.2503	85.178	1.694	0.070
15	876	937.5814	18.5443	-61.5814	85.250	-0.722	0.012
16	1016	995.8980	16.0475	20.1020	85.756	0.234	0.001
17	1044	1101	16.5925	-56.8680	85.652	-0.664	0.008
18	839	885.0964	21.9182	-46.0964	84.446	-0.546	0.010
19	1324	1392	38.6460	-68.4513	78.218	-0.875	0.093
20	937	990.0864	16.2192	-53.0864	85.723	-0.619	0.007
21	1178	995.8980	16.0475	182.1020	85.756	2.124	0.079
22	1170	990.0864	16.2192	179.9336	85.723	2.099	0.079
23	861	943.4130	18.2253	-82.4130	85.319	-0.966	0.021
24	1116	1130	17.8825	-14.0263	85.392	-0.164	0.001
25	1015	1008	15.7638	7.4386	85.808	0.087	0.000
26	888	955.0764	17.6284	-67.0764	85.445	-0.785	0.013
27	865	920.0864	19.5757	-55.0864	85.020	-0.648	0.011
28	1010	1002	15.8954	8.2703	85.784	0.096	0.000
29	929	949.2447	17.9197	-20.2447	85.384	-0.237	0.001
30	1053	1054	15.4888	-1.2147	85.858	-0.014	0.000
31	994	1008	15.7638	-13.5614	85.808	-0.158	0.000
32	1538	1655	63.7141	-116.8762	59.599	-1.961	2.198

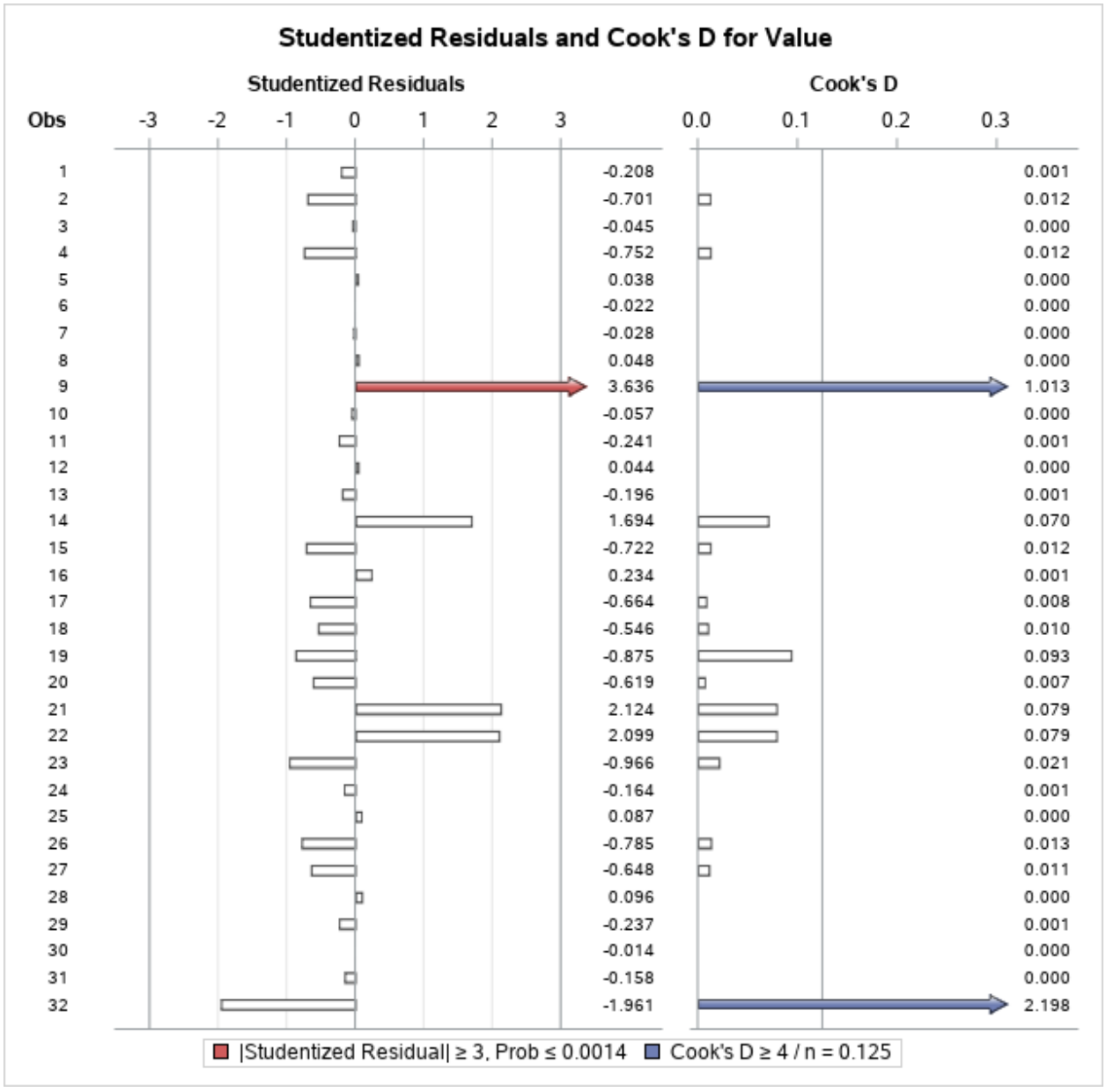
c. Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

By looking at the residual for the plot of the predicted values we can see that the majority of points are near zero residual lines which indicates a good model but few are far from it, therefore we need to test their influence by looking at the student residual and cook values.



The following table confirms the presence of two influential points (observations 9, 32) with cook values of 1.013 and 2.198 respectively. As well as observations 9 being an outlier.

With the assumption of linearity, normality, Independence, and Homoscedasticity of the data points. From there we conclude that our regression model is affected by two influential points and is not ideal with an accuracy of 0.7673.


Studentized Residuals and Cook's D for Value
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.0 0.1 0.2 0.3
1 -0.208 0.001
2 -0.701 0.012
3 -0.045 0.000
4 -0.752 0.012
5 0.038 0.000
6 -0.022 0.000
7 -0.028 0.000
8 0.048 0.000
9 3.636 1.013
10 -0.057 0.000
11 -0.241 0.001
12 0.044 0.000
13 -0.196 0.001
14 1.694 0.070
15 -0.722 0.012
16 0.234 0.001
17 -0.664 0.008
18 -0.546 0.010
19 -0.875 0.093
20 -0.619 0.007
21 2.124 0.079
22 2.099 0.079
23 -0.966 0.021
24 -0.164 0.001
25 0.087 0.000
26 -0.785 0.013
27 -0.648 0.011
28 0.096 0.000
29 -0.237 0.001
30 -0.014 0.000
31 -0.158 0.000
32 -1.961 2.198
|Studentized Residual| ≥ 3, Prob ≤ 0.0014
Cook's D ≥ 4 / n = 0.125