I.

A. For this first part, the goal is to predict the *Salary* variable.

a. Run a standard OLS regression using the variables just as they are (without *School* name), with no modifications at all, to predict the Salary variable. What is your fitted equation?

ANS:

Summary of Fit

RSquare	0.776441
RSquare Adj	0.763574
Root Mean Square Error	5732.184
Mean of Response	106169.4
Observations (or Sum Wgts)	148

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-2245118	1067756	-2.10	0.0373*	.
Rank	-239.372	159.3759	-1.50	0.1354	5.6994474
Year	1171.7303	544.4838	2.15	0.0331*	7.0066246
GPA	7092.1049	8351.545	0.85	0.3972	3.3100253
GMAT	-112.9887	64.36226	-1.76	0.0814	4.5280098
AcceptRate	-36530.63	9500.774	-3.85	0.0002*	3.0162129
EmpGrad	59202.693	5113.679	11.58	<.0001*	1.4211055
Tuition	0.527912	0.198646	2.66	0.0088*	6.6473814
Enrollment	2.5067514	1.585212	1.58	0.1161	1.6705831

$$Salary = -2245118 - 239 * Rank + 1171 * Year + 7092 * GPA - 112 * GMAT - 36530 * AcceptRate + 59202 * EmpGrad + 0.52 * Tuition + 2.50 * Enrollment$$

b. Are there any issues or violations with your output results from part a?

ANS:

Yes,

- Some independent variables are not statistically significant (at the usual 1%/5% threshold).

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-2245118	1067756	-2.10	0.0373*	.
Rank	-239.372	159.3759	-1.50	0.1354	5.6994474
Year	1171.7303	544.4838	2.15	0.0331*	7.0066246
GPA	7092.1049	8351.545	0.85	0.3972	3.3100253
GMAT	-112.9887	64.36226	-1.76	0.0814	4.5280098
AcceptRate	-36530.63	9500.774	-3.85	0.0002*	3.0162129
EmpGrad	59202.693	5113.679	11.58	<.0001*	1.4211055
Tuition	0.527912	0.198646	2.66	0.0088*	6.6473814
Enrollment	2.5067514	1.585212	1.58	0.1161	1.6705831

c. Attempt to run an FEM (Fixed Effects Model) with this data. (at this point, still don't do anything else and do NOT bother eliminating insignificant variables yet) Is the *group* of dummy variables statistically significant? Prove your answer statistically.

ANS:

Summary of Fit

RSquare	0.828468
RSquare Adj	0.791609
Root Mean Square Error	5381.607
Mean of Response	106169.4
Observations (or Sum Wgts)	148

To check statistically, calculate restricted F-test:

F(18,148-27) = (0.82 – 0.77)/18 / (1-0.82)/(148-27)

F(18,121) = (0.05/18) / (0.18/121) = 1.867

The F-Statistics at 5% level of significant, F(18,121), is 1.69 .

So, we can reject the null Hypothesis that all coefficients of dummy variables are 0 and conclude that the group of dummy variables is statistically significant.

d. Are there any other issues remaining in the output of the model from part c?

ANS:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-4771309	1600567	-2.98	0.0035*	.
Rank	760.38264	293.5515	2.59	0.0108*	21.936767
Year	2418.4029	808.6426	2.99	0.0034*	17.533499
GPA	17176.694	12817.59	1.34	0.1827	8.8455937
GMAT	-129.0005	81.77347	-1.58	0.1173	8.2925094
AcceptRate	-39183.6	11701.06	-3.35	0.0011*	5.1905224
EmpGrad	65864.942	5174.529	12.73	<.0001*	1.6508869
Tuition	0.0116323	0.318275	0.04	0.9709	19.360336
Enrollment	7.341597	9.685065	0.76	0.4499	70.748117
Carnegie Mellon	-3595.844	3285.67	-1.09	0.2760	2.8208527
Chicago	6894.071	6865.503	1.00	0.3173	7.8625821
Columbia	3693.9384	8017.147	0.46	0.6458	16.794695
Cornell	2660.2091	3564.954	0.75	0.4570	3.3207815
Dartmouth	8815.1116	3300.325	2.67	0.0086*	2.8460722
Duke	3877.7454	4338.507	0.89	0.3732	4.9182776
Georgetown	-11721.36	4497.191	-2.61	0.0103*	5.2846369
Harvard	2674.4184	13897.31	0.19	0.8477	50.465449
Michigan	3811.9403	5094.418	0.75	0.4558	6.7814333
MIT	8282.8113	4723.977	1.75	0.0821	5.8310665
Northwestern	5725.0115	7392.484	0.77	0.4402	12.583815
NYU	-131.3424	4523.149	-0.03	0.9769	5.3458183
Penn	6466.0903	12015.94	0.54	0.5915	37.72663
Stanford	11959.166	5115.794	2.34	0.0210*	6.8384637
UC Berkeley	591.40173	3861.963	0.15	0.8785	3.8971626
UCLA	-2795.441	4138.735	-0.68	0.5007	4.4757693
USC	-12409.56	3658.742	-3.39	0.0009*	3.4978084
UT Austin	-9438.389	4247.662	-2.22	0.0281*	4.7144641

- Some independent variables are highly correlated with others (high VIF) or not statistically significant (at the usual 1%/5% threshold).

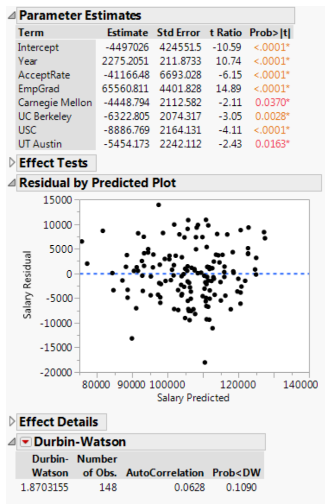
e. Work with the data to correct for the problem you noted in part d. (You shouldn't need to spend TOO much time doing this!) Again, predict the *Salary* variable. Include a brief report of any models you tried and why you discarded them. (maybe a short table) What is your final model, including the output and residual plot?

ANS:

Step	Explanation
Drop Enrollment	Too high VIF(70)
Drop Rank	Too high VIF(22)
Drop Tuition	Not significant at 5%
Drop UCLA	Not significant at 5%
Drop NYU	Not significant at 5%
Drop GPA	Not significant at 5%
Drop Cornell	Not significant at 5%
Drop Michigan	Not significant at 5%
Drop Duke	Not significant at 5%
Drop Northwestern	Not significant at 5%
Drop Dartmouth	Not significant at 5%

Drop MIT	Not significant at 5%
Drop Columbia	Not significant at 5%
Drop Chicago	Not significant at 5%
Drop Harvard	Not significant at 5%
Drop Stanford	Not significant at 5%
Drop GMAT	Not significant at 5%
Drop Georgetown	Not significant at 5%
Drop Penn	Not significant at 5%

This is the output of my final model:


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4497026	424551.5	-10.59	<.0001*
Year	2275.2051	211.8733	10.74	<.0001*
AcceptRate	-41166.48	6693.028	-6.15	<.0001*
EmpGrad	65560.811	4401.828	14.89	<.0001*
Carnegie Mellon	-4448.794	2112.582	-2.11	0.0370*
UC Berkeley	-6322.805	2074.317	-3.05	0.0028*
USC	-8886.769	2164.131	-4.11	<.0001*
UT Austin	-5454.173	2242.112	-2.43	0.0163*

Effect Tests

Residual by Predicted Plot

Effect Details

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.8703155	148	0.0628	0.1090

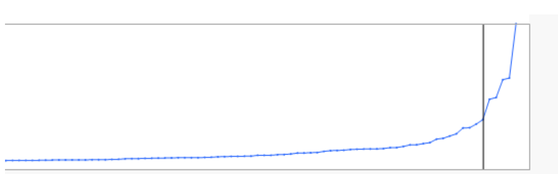
The final model is:

$$Salary = -4497026 + 2275 * Year - 41166 * AcceptRate + 65560 * EmpGrad - 4448 * CMU - 6322 * UC\_Berkeley - 8886 * USC - 5454 * UT\_AUSTIN$$

B. For this second part, you are to use your knowledge of Clustering Analysis.

a. Using a combination of *Hierarchical* and *K-Means* clustering methods, what number of clusters would you suggest using all the data except the *School* names? What is your reason for choosing that number?

<u>ANS:</u>



14	3.49046925	16	111
13	3.94964355	5	7
12	4.05859894	36	43
11	4.35948275	3	53
10	4.64580625	17	19
9	5.41505939	1	9
8	5.44175397	11	113
7	5.92851421	16	61
6	6.50578150	11	36
5	9.16776132	3	5
4	9.39586083	11	17
3	11.72526904	1	3
2	11.94021095	11	16
1	19.03502322	1	11

As the number of observations is not too big, I use Hierarchical clustering for determining the optimal number of group. The scree plot shows that the Distance slowly drops after splitting into 6 clusters. Based on this observation, the data should be split into 6 groups.

b. Regardless of your answer to the previous question, form 6 clusters of the data and save the cluster numbers to your worksheet (under the red triangle, choose *Save Clusters*). Create a summary table of the clusters' statistics and copy it to your answer sheet. (*Analyze – Tabulate,* drag summary statistics of *N, Mean, Std Dev, Min,* and *Max* to the top to be the column headings. Then drag the *Cluster Number* variable to the left side to be the row headings, and finally select the 8 numerical variables and drag them just to the right of the *Cluster Number* variable. You can save this to its own table by using the red triangle next to *Tabulate*.)

ANS:

c. How would you characterize each cluster? Give just a brief description of each of the 6 in a summary table.

ANS:

d. Which, if any, of the quantitative variables are significantly predictable using only the *Cluster Number?* Justify your answer statistically. What method did you use to answer this?

ANS:

II.

a. What type of analysis is appropriate to analyze this? What are the null hypotheses for these two goals?

ANS:

ANOVA Analysis

- There is no different in the means of US Domestic Revenues between different movie Genres

- There is no different in the means of US Domestic Revenues between different MPAA Ratings

b. Determine whether there is a statistically significant difference overall in *US Domestic Revenues* between different *Genres*. Which *Genres*, if any, are statistically different from each other in terms of *US Domestic Revenues*?

ANS:

Calculate F-statistics from ANOVA:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Genre	10	245966.66	24596.7	3.3251	0.0016*
Error	64	473420.90	7397.2		
C. Total	74	719387.56			

At 5% significant level, we can conclude that there is a significant difference overall in US Domestic Revenues between different Genres.

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
thriller	comedy	238.4333	99.3123	40.034	436.8325	0.0193*
fantasy	comedy	235.0833	65.6889	103.855	366.3120	0.0007*
thriller	biography	215.6000	90.2049	35.395	395.8050	0.0198*
fantasy	biography	212.2500	50.8824	110.601	313.8993	<.0001*
thriller	drama	209.0465	87.0013	35.241	382.8516	0.0192*
fantasy	drama	205.6965	44.9592	115.880	295.5128	<.0001*
thriller	musical	179.4500	105.3366	-30.984	389.8841	0.0933
fantasy	musical	176.1000	74.4842	27.301	324.8994	0.0211*
thriller	war	167.2500	105.3366	-43.184	377.6841	0.1173
fantasy	war	163.9000	74.4842	15.101	312.6994	0.0314*
thriller	crime	162.8000	121.6322	-80.188	405.7883	0.1855
fantasy	crime	159.4500	96.1587	-32.649	351.5492	0.1022
thriller	western	150.8000	105.3366	-59.634	361.2341	0.1571
thriller	animation	147.6000	121.6322	-95.388	390.5883	0.2294
fantasy	western	147.4500	74.4842	-1.349	296.2494	0.0521
fantasy	animation	144.2500	96.1587	-47.849	336.3492	0.1385
romance	comedy	126.5167	60.8161	5.022	248.0108	0.0415*
thriller	romance	111.9167	92.8982	-73.669	297.5021	0.2327
fantasy	romance	108.5667	55.5173	-2.342	219.4752	0.0549
romance	biography	103.6833	44.4138	14.957	192.4101	0.0227*
romance	drama	97.1298	37.4819	22.251	172.0086	0.0118*
animation	comedy	90.8333	99.3123	-107.566	289.2325	0.3638

From the above data, significantly difference pairs of Genres are:

Genre1	Genre2
Thriller	Comedy
Fantasy	Comedy
Thriller	Biography
Fantasy	Biography
Thriller	Drama
Fantasy	Drama
Fantasy	Musical
Fantasy	War
Romance	Comedy
Romance	Biography
Romance	Drama

c. Determine whether there is a statistically significant difference overall in *US Domestic Revenues* between different *MPAA Ratings*. Which *Ratings*, if any, are statistically different from each other in terms of *US Domestic Revenues*?

ANS:

Calculate F-statistics from ANOVA:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
MPAA rating	3	126864.81	42288.3	5.0673	0.0031*
Error	71	592522.75	8345.4		
C. Total	74	719387.56			

At 5% significant level, we can conclude that there is a significant difference overall in US Domestic Revenues between different MPAA Ratings.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
PG-13	PG	92.24857	39.06423	14.357	170.1404	0.0210*
PG-13	R	87.02732	23.18106	40.806	133.2490	0.0004*
PG-13	G	60.37000	67.13055	-73.485	194.2245	0.3715

From the above data, significantly difference pairs of MPAA Ratings are:

MPAA Rating1	MPAA Rating2
PG-13	PG
PG-13	R

d. Give a *general* interpretation of your final analysis.

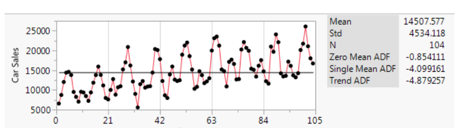
<u>ANS:</u>

- In general, there is a significant difference overall in US Domestic Revenues between different Genres. However, a different in some pairs of Genres is not conclusive and all Genres do not have completely different US Domestic Revenues.
- US Domestic Revenues of PG-13 films are significantly different from PG and R films' US Domestic Revenues. The different in US Domestic Revenues for other pairs of MPAA Rating are not so apparent.

III.

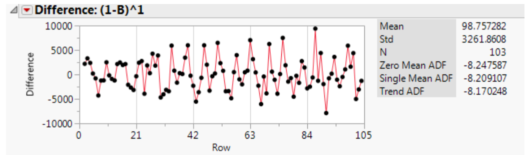
a. Is the *Car Sales* variable stationary? How do you know (statistically)? If it is *not* stationary, what needs to be done to make it stationary? Justify your answer

<u>ANS:</u>



Car Sales is not stationary, as shown by the Dickey-Fuller test of no constant term (Zero Mean ADF). It does not pass at either 1% or 5% significant threshold.

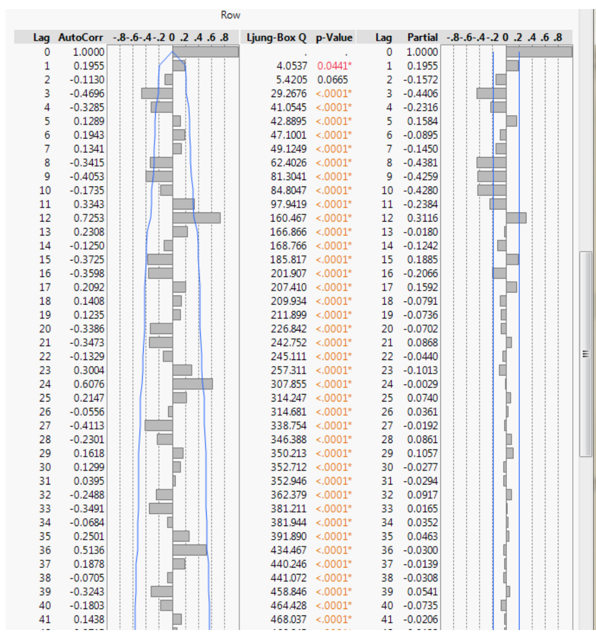
To make it stationary, take a difference of 1 lag of the series. Now, all Dickey-Fuller tests indicate that the series is stationary at 1% significant level.



b. Does there seem to be a seasonal, or cyclical, pattern to the data? Justify your answer

ANS:

Yes, continue from a., since the data come from monthly observation, we look at the significant spikes for ACF and PACF plots for every 12 lags. ACF plot suggests that up to 3 periods of seasonal MA term should be included. PACF plot suggests that up to 1 period of seasonal AR term should be included.



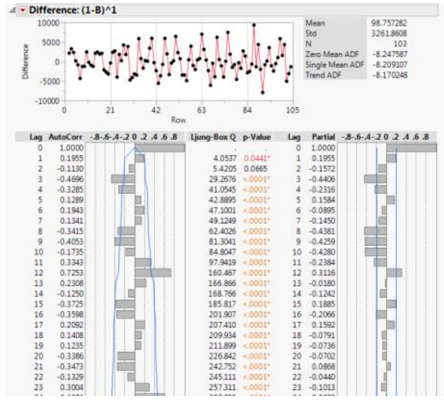
Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.1955	4.0537	0.0441*	1	0.1955
2	-0.1130	5.4205	0.0665	2	-0.1572
3	-0.4696	29.2676	<.0001*	3	-0.4406
4	-0.3285	41.0545	<.0001*	4	-0.2316
5	0.1289	42.8895	<.0001*	5	0.1584
6	0.1943	47.1001	<.0001*	6	-0.0895
7	0.1341	49.1249	<.0001*	7	-0.1450
8	-0.3415	62.4026	<.0001*	8	-0.4381
9	-0.4053	81.3041	<.0001*	9	-0.4259
10	-0.1735	84.8047	<.0001*	10	-0.4280
11	0.3343	97.9419	<.0001*	11	-0.2384
12	0.7253	160.467	<.0001*	12	0.3116
13	0.2308	166.866	<.0001*	13	-0.0180
14	-0.1250	168.766	<.0001*	14	-0.1242
15	-0.3725	185.817	<.0001*	15	0.1885
16	-0.3598	201.907	<.0001*	16	-0.2066
17	0.2092	207.410	<.0001*	17	0.1592
18	0.1408	209.934	<.0001*	18	-0.0791
19	0.1235	211.899	<.0001*	19	-0.0736
20	-0.3386	226.842	<.0001*	20	-0.0702
21	-0.3473	242.752	<.0001*	21	0.0868
22	-0.1329	245.111	<.0001*	22	-0.0440
23	0.3004	257.311	<.0001*	23	-0.1013
24	0.6076	307.855	<.0001*	24	-0.0029
25	0.2147	314.247	<.0001*	25	0.0740
26	-0.0556	314.681	<.0001*	26	0.0361
27	-0.4113	338.754	<.0001*	27	-0.0192
28	-0.2301	346.388	<.0001*	28	0.0861
29	0.1618	350.213	<.0001*	29	0.1057
30	0.1299	352.712	<.0001*	30	-0.0277
31	0.0395	352.946	<.0001*	31	-0.0294
32	-0.2488	362.379	<.0001*	32	0.0917
33	-0.3491	381.211	<.0001*	33	0.0165
34	-0.0684	381.944	<.0001*	34	0.0352
35	0.2501	391.890	<.0001*	35	0.0463
36	0.5136	434.467	<.0001*	36	-0.0300
37	0.1878	440.246	<.0001*	37	-0.0139
38	-0.0705	441.072	<.0001*	38	-0.0308
39	-0.3243	458.846	<.0001*	39	0.0541
40	-0.1803	464.428	<.0001*	40	-0.0735
41	0.1438	468.037	<.0001*	41	-0.0206

c. Using your vast knowledge of *Time Series* analysis, create a solid model to predict *Car Sales* for the next 4 months (I actually have the true values and can compare them to your estimates!), including 95% intervals.

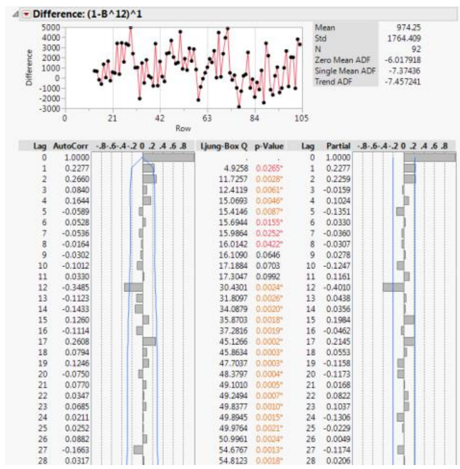
ANS:

I look at the 3 versions of differences:

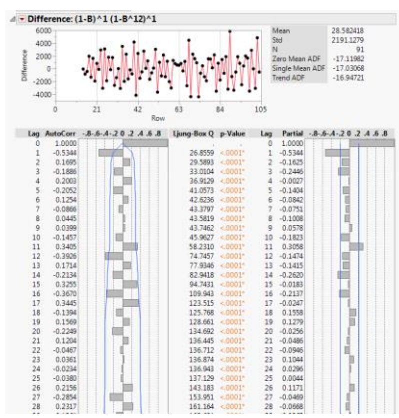
1.) Non-Seasonal 1st difference



2.) Seasonal 1st difference

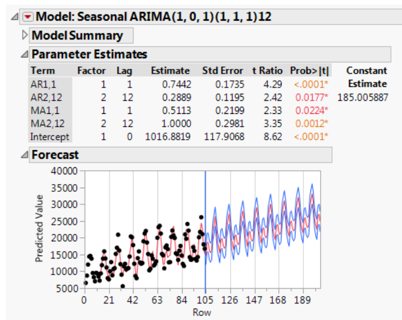


3.) Both non-seasonal and seasonal 1st difference

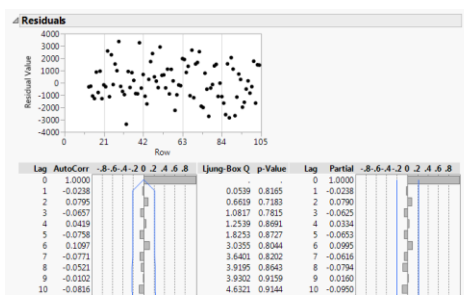


All are statistically stationary but it is the 2nd version (1st difference on just seasonal part) that yields the most compelling plot (ACF and PACF plot die down pretty fast and not too much random spikes). So, I decided to run the ARIMA group model based on the suggestion of the 2nd version:

Among the top models, I select the 2nd model, ARIMA(1,0,1)x(1,1,1), because all parameters are significant(at 5% threshold), prediction error, r-square and AIC are not so much different from the top model.



I don't see any problems in the residual plot and it appears to be stationary.



This is the prediction of Car Sales for the next 4 months:

Prediction	Upper 95%	Lower 95%
14753.447665	17445.223906	12061.671424
18576.585669	21340.341807	15812.82953
18579.219469	21382.040037	15776.398902
16254.336129	19078.556556	13430.115702