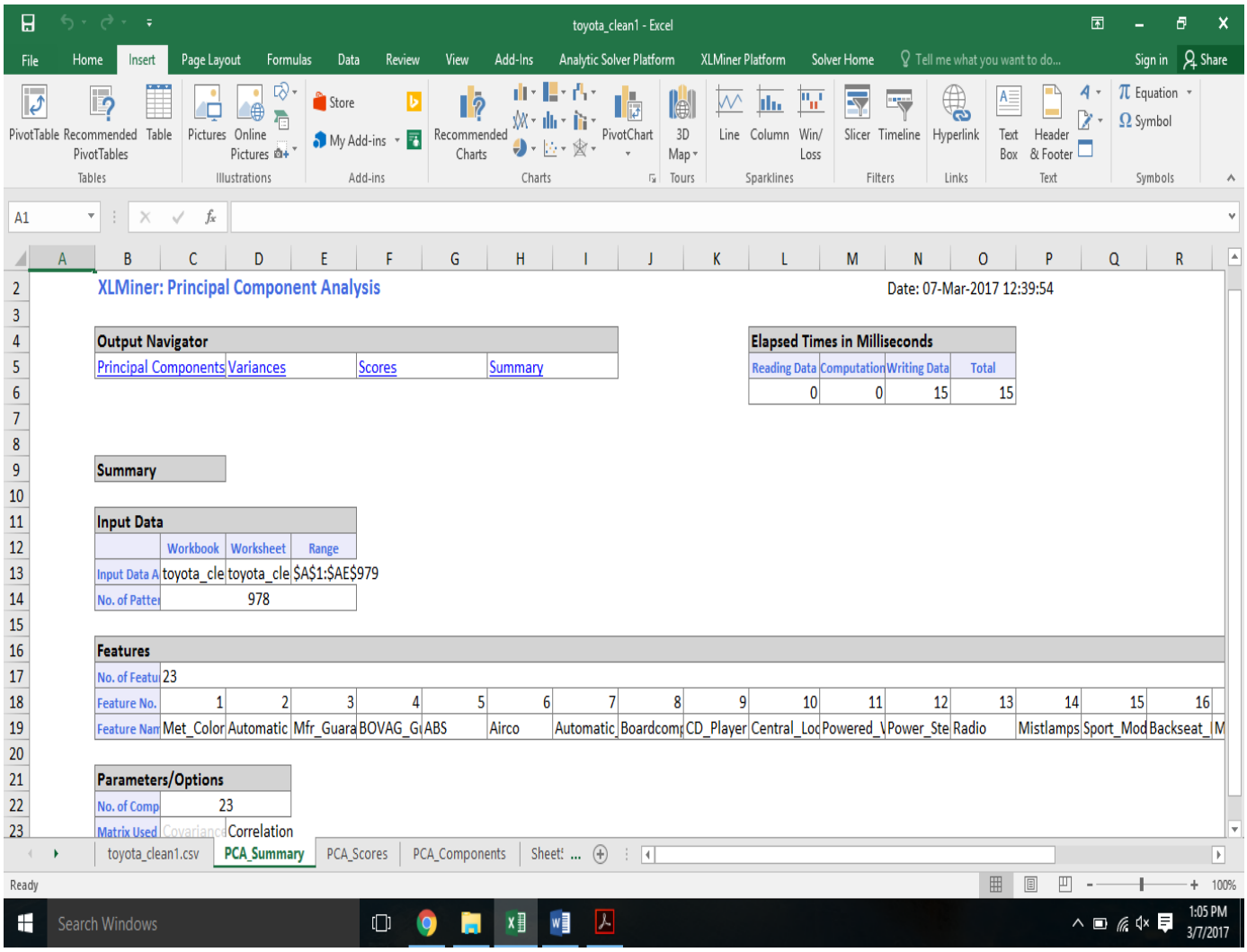
Archit Jajoo
ISDS 574

HOMEWORK 1

a) After removing the reference variable which is 'Grey' we run the correlation based-PCA and get the following:



XLMiner: Principal Component Analysis

Date: 07-Mar-2017 12:39:54

Output Navigator

Principal Components	Variances	Scores	Summary

Elapsed Times in Milliseconds

Reading Data	Computation	Writing Data	Total
0	0	15	15

Summary

Input Data

	Workbook	Worksheet	Range
Input Data A	toyota_cle	toyota_cle	$A$1:$AE$979
No. of Patte	978		

Features

No. of Featu	23															
Feature No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Feature Nam	Met_Color	Automatic	Mfr_Guara	BOVAG_G	ABS	Airco	Automatic_	Boardcom	CD_Player	Central_Loc	Powered_\	Power_Ste	Radio	Mistlamps	Sport_Mod	Backseat_M

Parameters/Options

No. of Comp	23
Matrix Used	Covariance / Correlation

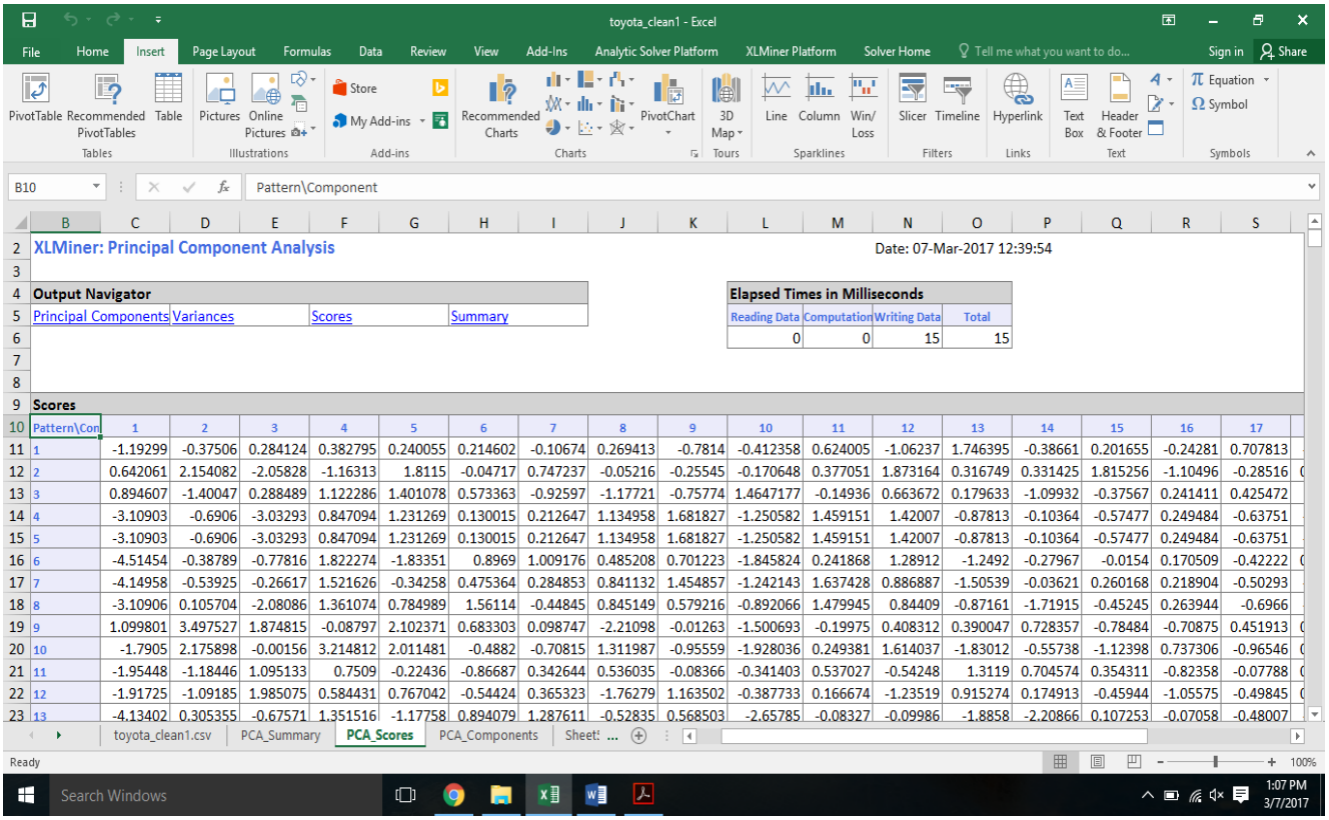
toyota_clean1.csv | PCA_Summary | PCA_Scores | PCA_Components | Sheet1 ...

Ready

Search Windows

1:05 PM
3/7/2017

Archit Jajoo
ISDS 574



toyota_clean1 - Excel

B10: Pattern\Component

XLMiner: Principal Component Analysis

Date: 07-Mar-2017 12:39:54

Output Navigator

Principal Components | Variances | Scores | Summary

Elapsed Times in Milliseconds

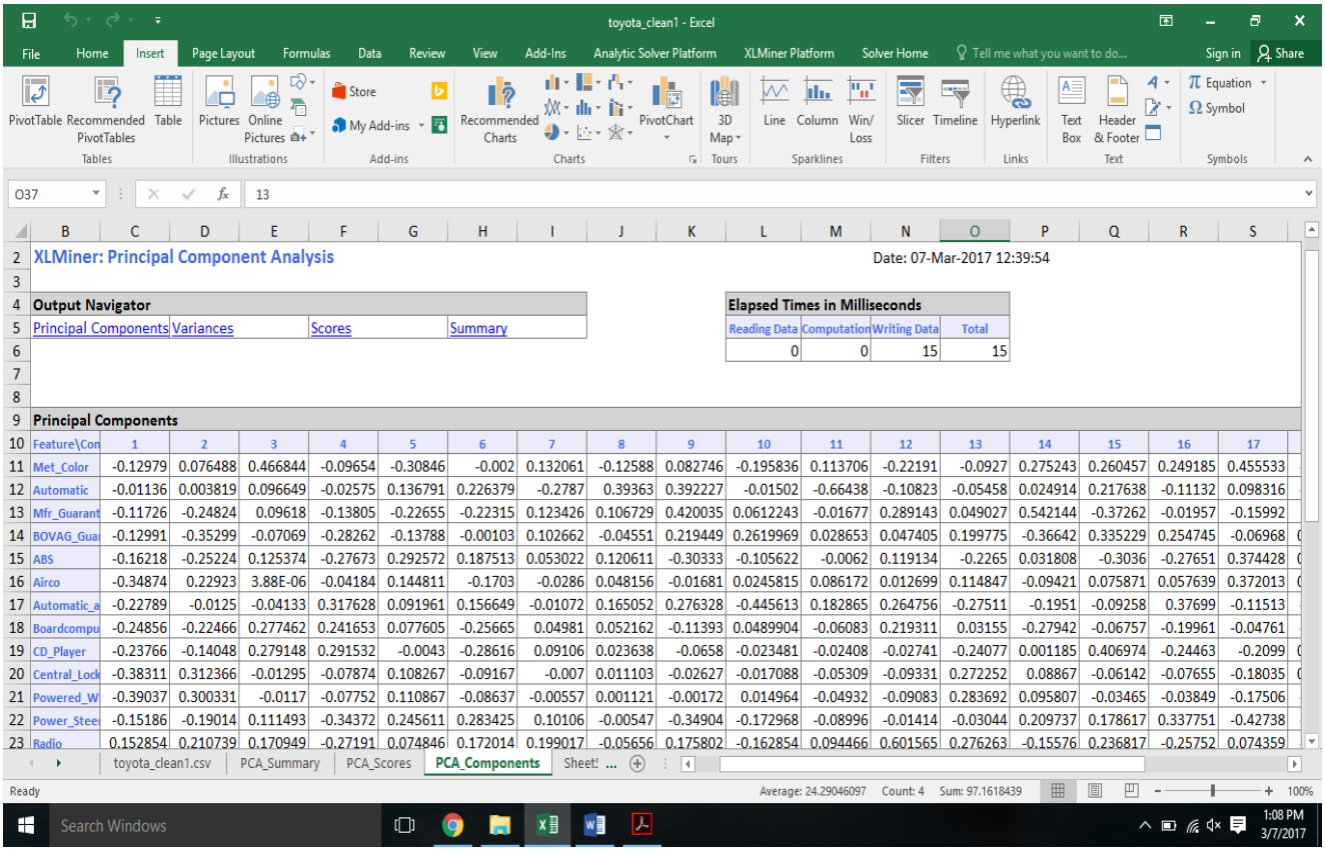
Reading Data	Computation	Writing Data	Total
0	0	15	15

Scores

Pattern\Com	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	-1.19299	-0.37506	0.284124	0.382795	0.240055	0.214602	-0.10674	0.269413	-0.7814	-0.412358	0.624005	-1.06237	1.746395	-0.38661	0.201655	-0.24281	0.707813
2	0.642061	2.154082	-2.05828	-1.16313	1.8115	-0.04717	0.747237	-0.05216	-0.25545	-0.170648	0.377051	1.873164	0.316749	0.331425	1.815256	-1.10496	-0.28516
3	0.894607	-1.40047	0.288489	1.122286	1.401078	0.573363	-0.92597	-1.17721	-0.75774	1.4647177	-0.14936	0.663672	0.179633	-1.09932	-0.37567	0.241411	0.425472
4	-3.10903	-0.6906	-3.03293	0.847094	1.231269	0.130015	0.212647	1.134958	1.681827	-1.250582	1.459151	1.42007	-0.87813	-0.10364	-0.57477	0.249484	-0.63751
5	-3.10903	-0.6906	-3.03293	0.847094	1.231269	0.130015	0.212647	1.134958	1.681827	-1.250582	1.459151	1.42007	-0.87813	-0.10364	-0.57477	0.249484	-0.63751
6	-4.51454	-0.38789	-0.77816	1.822274	-1.83351	0.8969	1.009176	0.485208	0.701223	-1.845824	0.241868	1.28912	-1.2492	-0.27967	-0.0154	0.170509	-0.42222
7	-4.14958	-0.53925	-0.26617	1.521626	-0.34258	0.475364	0.284853	0.841132	1.454857	-1.242143	1.637428	0.886887	-1.50539	-0.03621	0.260168	0.218904	-0.50293
8	-3.10906	0.105704	-2.08086	1.361074	0.784989	1.56114	-0.44845	0.845149	0.579216	-0.892066	1.479945	0.84409	-0.87161	-1.71915	-0.45245	0.263944	-0.6966
9	1.099801	3.497527	1.874815	-0.08797	2.102371	0.683303	0.098747	-2.21098	-0.01263	-1.500693	-0.19975	0.408312	0.390047	0.728357	-0.78484	-0.70875	0.451913
10	-1.7905	2.175898	-0.00156	3.214812	2.011481	-0.4882	-0.70815	1.311987	-0.95559	-1.928036	0.249381	1.614037	-1.83012	-0.55738	-1.12398	0.737306	-0.96546
11	-1.95448	-1.18446	1.095133	0.7509	-0.22436	-0.86687	0.342644	0.536035	-0.08366	-0.341403	0.537027	-0.54248	1.3119	0.704574	0.354311	-0.82358	-0.07788
12	-1.91725	-1.09185	1.985075	0.584431	0.767042	-0.54424	0.365323	-1.76279	1.163502	-0.387733	0.166674	-1.23519	0.915274	0.174913	-0.45944	-1.05575	-0.49845
13	-4.13402	0.305355	-0.67571	1.351516	-1.17758	0.894079	1.287611	-0.52835	0.568503	-2.65785	-0.08327	-0.09986	-1.8858	-2.20866	0.107253	-0.07058	-0.48007

toyota_clean1.csv | PCA_Summary | PCA_Scores | PCA_Components | Sheet5

Ready



toyota_clean1 - Excel

O37: 13

XLMiner: Principal Component Analysis

Date: 07-Mar-2017 12:39:54

Output Navigator

Principal Components | Variances | Scores | Summary

Elapsed Times in Milliseconds

Reading Data	Computation	Writing Data	Total
0	0	15	15

Principal Components

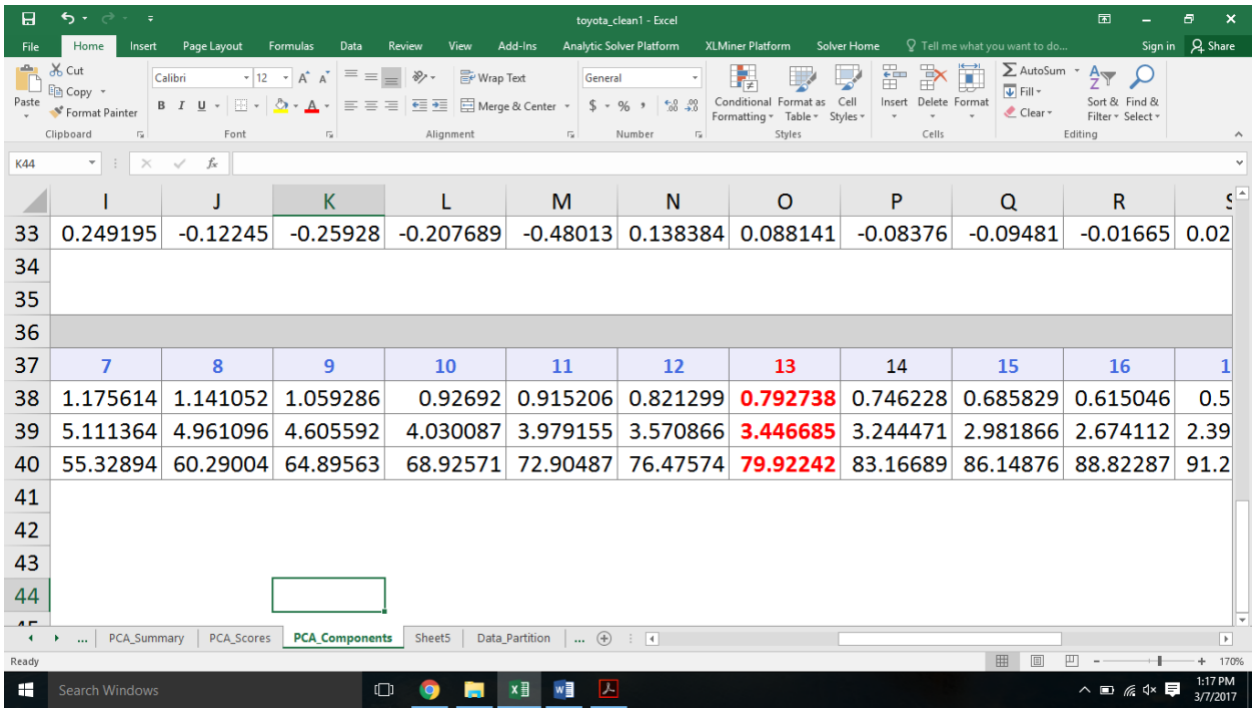
Feature\Com	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Met_Color	-0.12979	0.076488	0.466844	-0.09654	-0.30846	-0.002	0.132061	-0.12588	0.082746	-0.195836	0.113706	-0.22191	-0.0927	0.275243	0.260457	0.249185	0.455533
Automatic	-0.01136	0.003819	0.096649	-0.02575	0.136791	0.226379	-0.2787	0.39363	0.392227	-0.01502	-0.66438	-0.10823	-0.05458	0.024914	0.217638	-0.11132	0.098316
Mfr_Guarant	-0.11726	-0.24824	0.09618	-0.13805	-0.22655	-0.22315	0.123426	0.106729	0.420035	0.0612243	-0.01677	0.289143	0.049027	0.542144	-0.37262	-0.01957	-0.15992
BOVAG_Gua	-0.12991	-0.35299	-0.07069	-0.28262	-0.13788	-0.00103	0.102662	-0.04551	0.219449	0.2619969	0.028653	0.047405	0.199775	-0.36642	0.335229	0.254745	-0.06968
ABS	-0.16218	-0.25224	0.125374	-0.27673	0.292572	0.187513	0.053022	0.120611	-0.30333	-0.105622	-0.0062	0.119134	-0.2265	0.031808	-0.3036	-0.27651	0.374428
Airco	-0.34874	0.22923	3.88E-06	-0.04184	0.144811	-0.1703	-0.0286	0.048156	-0.01681	0.0245815	0.086172	0.012699	0.114847	-0.09421	0.075871	0.057639	0.372013
Automatic_a	-0.22789	-0.0125	-0.04133	0.317628	0.091961	0.156649	-0.01072	0.165052	0.276328	-0.445613	0.182865	0.264756	-0.27511	-0.1951	-0.09258	0.37699	-0.11513
Boardcompu	-0.24856	-0.22466	0.277462	0.241653	0.077605	-0.25665	0.04981	0.052162	-0.11393	0.0489904	-0.06083	0.219311	0.03155	-0.27942	-0.06757	-0.19961	-0.04761
CD_Player	-0.23766	-0.14048	0.279148	0.291532	-0.0043	-0.28616	0.09106	0.023638	-0.0658	-0.023481	-0.02408	-0.02741	-0.24077	0.001185	0.406974	-0.24463	-0.2099
Central_Lock	-0.38311	0.312366	-0.01295	-0.07874	0.108267	-0.09167	-0.007	0.011103	-0.02627	-0.017088	-0.05309	-0.09331	0.272252	0.08867	-0.06142	-0.07655	-0.18035
Powered_W	-0.39037	0.300331	-0.0117	-0.07752	0.110867	-0.08637	-0.00557	0.001121	-0.00172	0.014964	-0.04932	-0.09083	0.283692	0.095807	-0.03465	-0.03849	-0.17506
Power_Stee	-0.15186	-0.19014	0.111493	-0.34372	0.245611	0.283425	0.10106	-0.00547	-0.34904	-0.172968	-0.08996	-0.01414	-0.03044	0.209737	0.178617	0.337751	-0.42738
Radio	0.152854	0.210739	0.170949	-0.27191	0.074846	0.172014	0.199017	-0.05656	0.175802	-0.162854	0.094466	0.601565	0.276263	-0.15576	0.236817	-0.25752	0.074359

toyota_clean1.csv | PCA_Summary | PCA_Scores | PCA_Components | Sheet5

Ready

Average: 24.29046097 Count: 4 Sum: 97.1618439

b) We should keep 13 PC to capture at least 80% of total variance.



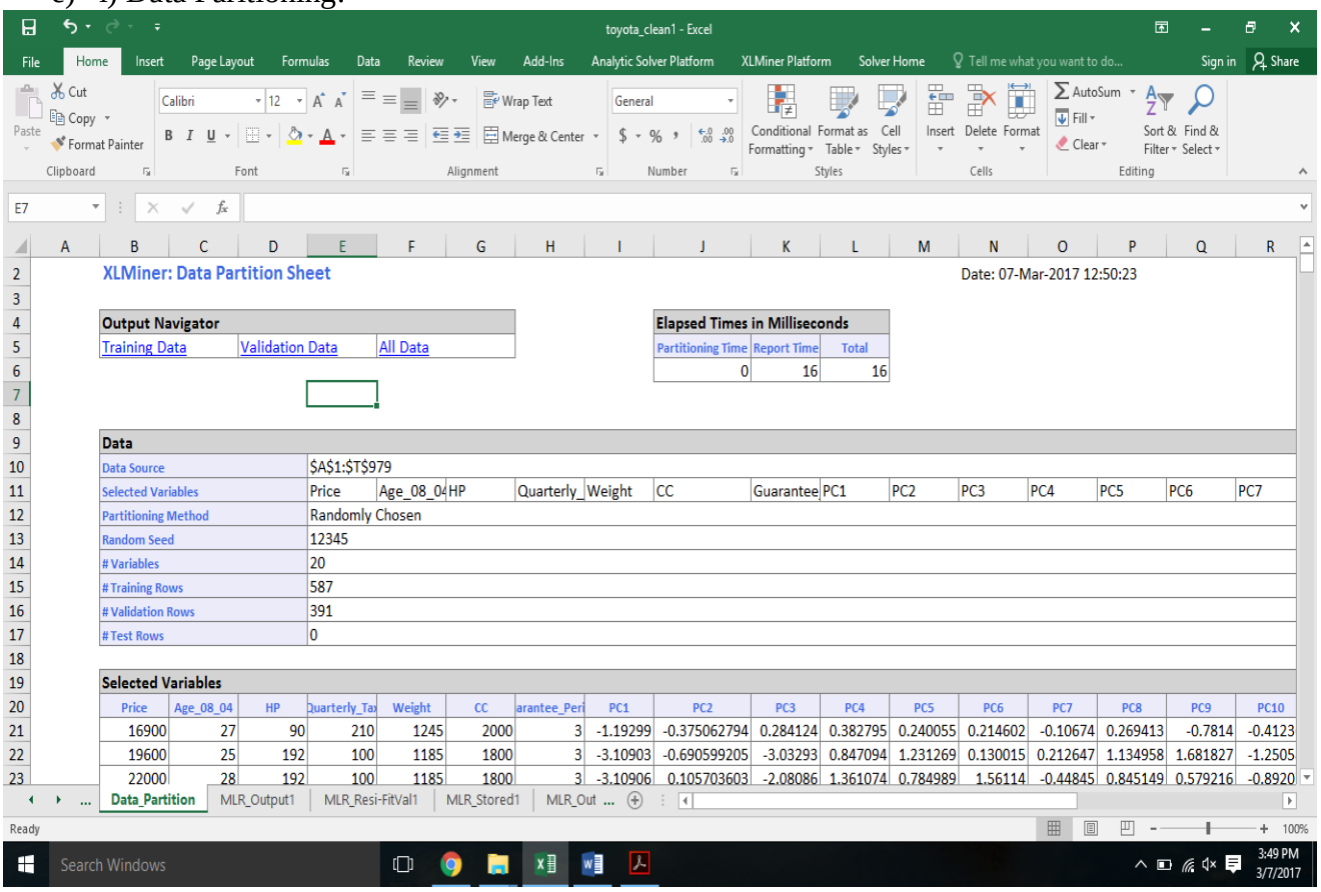
	I	J	K	L	M	N	O	P	Q	R	S
33	0.249195	-0.12245	-0.25928	-0.207689	-0.48013	0.138384	0.088141	-0.08376	-0.09481	-0.01665	0.02
34											
35											
36											
37	7	8	9	10	11	12	13	14	15	16	1
38	1.175614	1.141052	1.059286	0.92692	0.915206	0.821299	0.792738	0.746228	0.685829	0.615046	0.5
39	5.111364	4.961096	4.605592	4.030087	3.979155	3.570866	3.446685	3.244471	2.981866	2.674112	2.39
40	55.32894	60.29004	64.89563	68.92571	72.90487	76.47574	79.92242	83.16689	86.14876	88.82287	91.2
41											
42											
43											
44											

c) Principal Components are small number of uncorrelated variables taken from a large set of data and are identified through Principal Component Analysis.
The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score).

d) The variables which contribute the the first principal component are:
1. Powered_Windows = -0.390365139
2. Central_Lock = -0.383106122
3. Airco = -0.348743471

The Principal components are always selected in their highest order i.e. selection starts from the best.

e) i) Data Paritioning:



XLMiner: Data Partition Sheet

Date: 07-Mar-2017 12:50:23

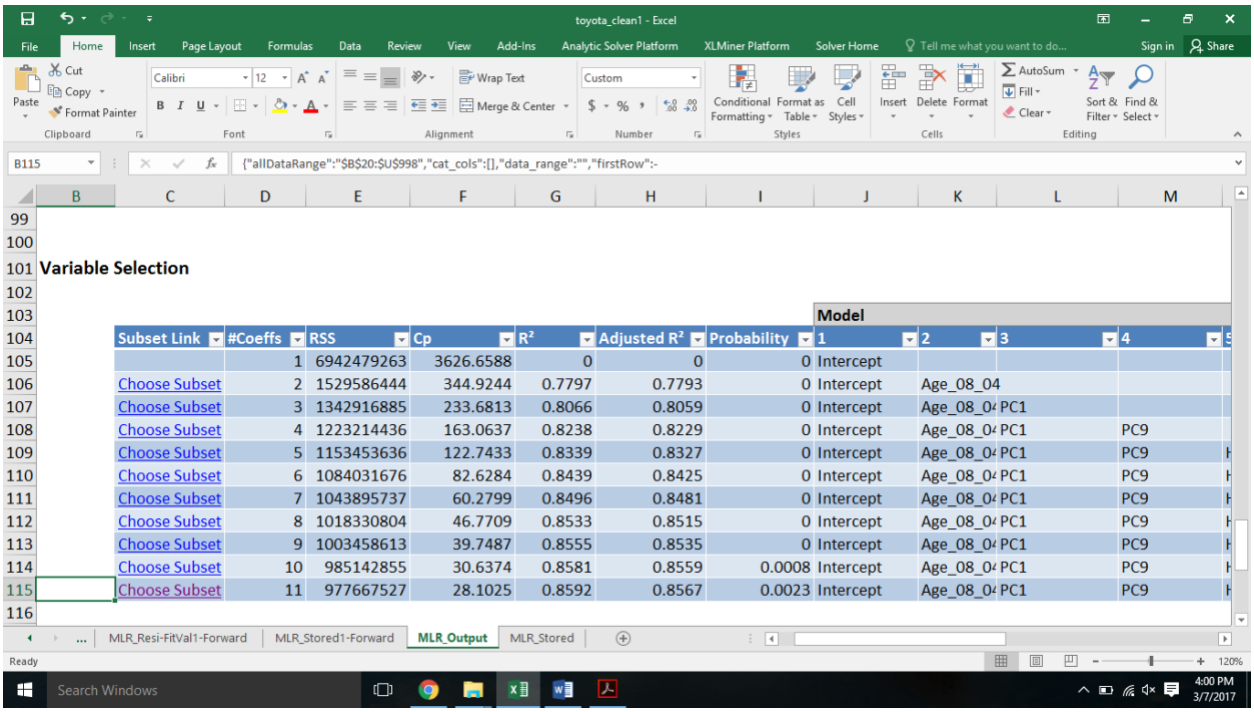
Output Navigator		
Training Data	Validation Data	All Data

Elapsed Times in Milliseconds		
Partitioning Time	Report Time	Total
0	16	16

Data	
Data Source	$A$1:$T$979
Selected Variables	Price, Age_08_04, HP, Quarterly_Tax, Weight, CC, Guarantee_Period, PC1, PC2, PC3, PC4, PC5, PC6, PC7
Partitioning Method	Randomly Chosen
Random Seed	12345
# Variables	20
# Training Rows	587
# Validation Rows	391
# Test Rows	0

Selected Variables

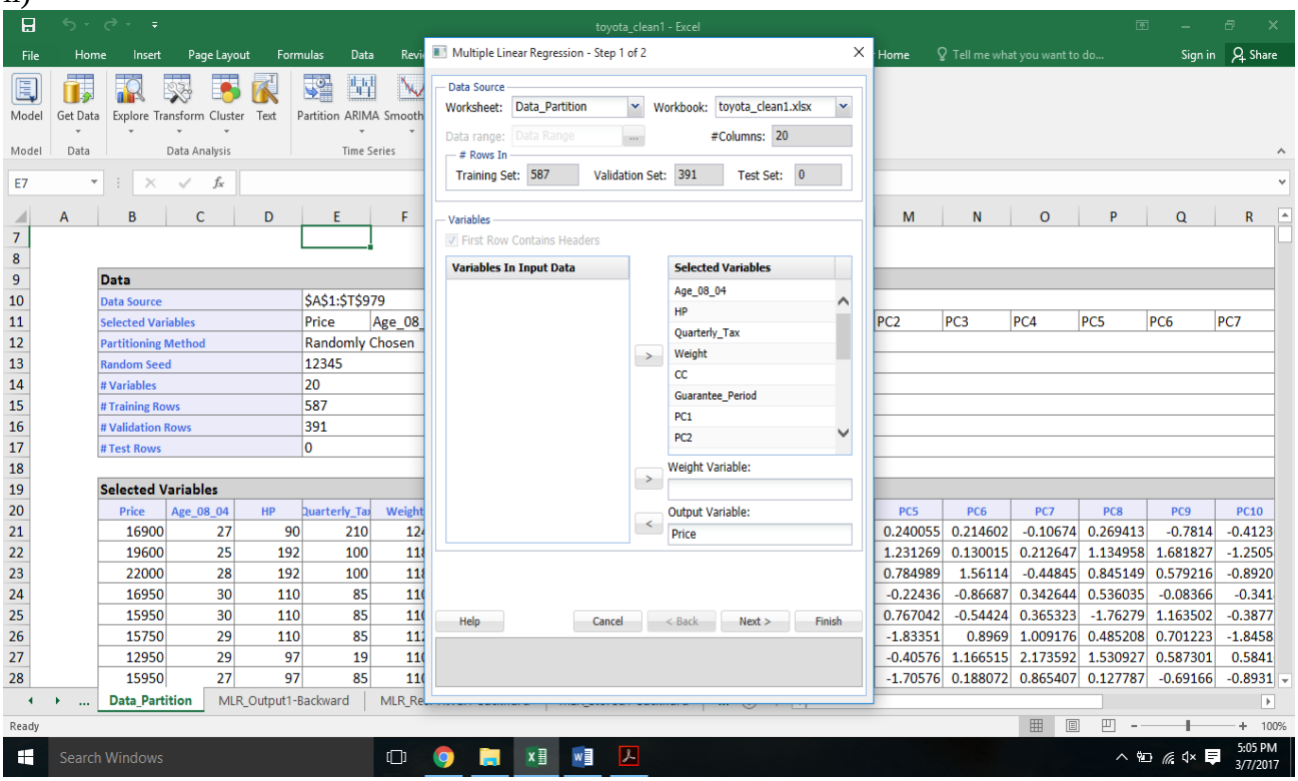
Price	Age_08_04	HP	Quarterly_Tax	Weight	CC	Guarantee_Period	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
16900	27	90	210	1245	2000	3	-1.19299	-0.375062794	0.284124	0.382795	0.240055	0.214602	-0.10674	0.269413	-0.7814	-0.4123
19600	25	192	100	1185	1800	3	-3.10903	-0.690599205	-3.03293	0.847094	1.231269	0.130015	0.212647	1.134958	1.681827	-1.2505
22000	28	192	100	1185	1800	3	-3.10906	0.105703603	-2.08086	1.361074	0.784989	1.56114	-0.44845	0.845149	0.579216	-0.8920



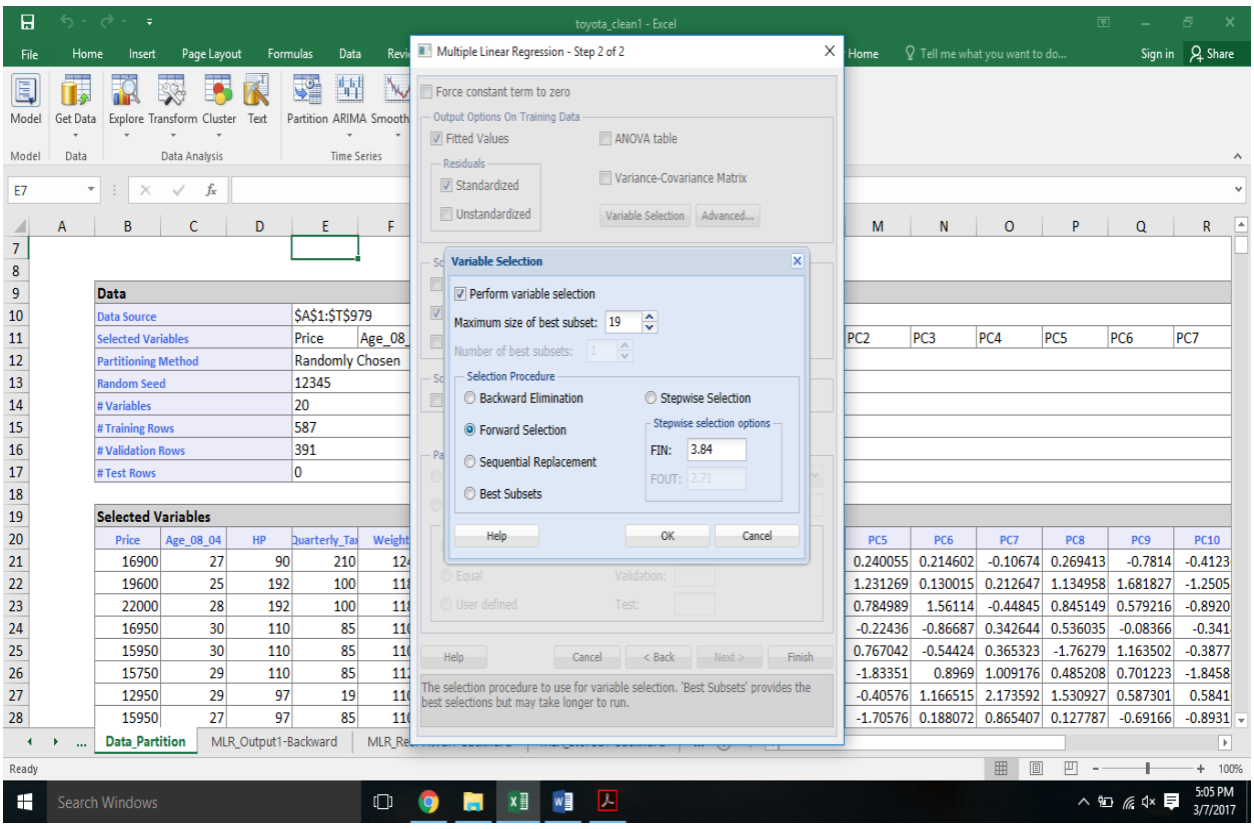
Variable Selection

							Model			
Subset Link	#Coeffs	RSS	Cp	R^2	Adjusted R^2	Probability	1	2	3	4
	1	6942479263	3626.6588	0	0	0	Intercept			
Choose Subset	2	1529586444	344.9244	0.7797	0.7793	0	Intercept	Age_08_04		
Choose Subset	3	1342916885	233.6813	0.8066	0.8059	0	Intercept	Age_08_04	PC1	
Choose Subset	4	1223214436	163.0637	0.8238	0.8229	0	Intercept	Age_08_04	PC1	PC9
Choose Subset	5	1153453636	122.7433	0.8339	0.8327	0	Intercept	Age_08_04	PC1	PC9
Choose Subset	6	1084031676	82.6284	0.8439	0.8425	0	Intercept	Age_08_04	PC1	PC9
Choose Subset	7	1043895737	60.2799	0.8496	0.8481	0	Intercept	Age_08_04	PC1	PC9
Choose Subset	8	1018330804	46.7709	0.8533	0.8515	0	Intercept	Age_08_04	PC1	PC9
Choose Subset	9	1003458613	39.7487	0.8555	0.8535	0	Intercept	Age_08_04	PC1	PC9
Choose Subset	10	985142855	30.6374	0.8581	0.8559	0.0008	Intercept	Age_08_04	PC1	PC9
Choose Subset	11	977667527	28.1025	0.8592	0.8567	0.0023	Intercept	Age_08_04	PC1	PC9

ii)

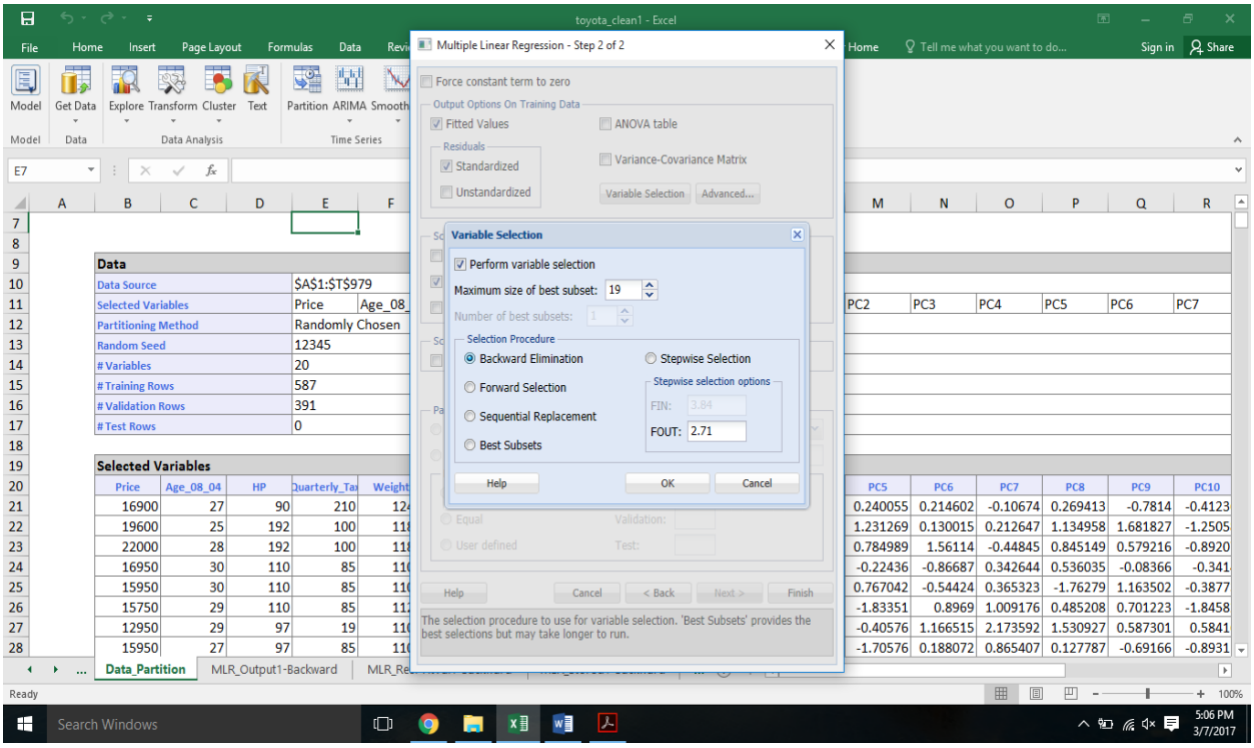

toyota_clean1 - Excel
Multiple Linear Regression - Step 1 of 2
Data Source
Worksheet: Data_Partition
Workbook: toyota_clean1.xlsx
Data range: Data Range
#Columns: 20
Rows In
Training Set: 587
Validation Set: 391
Test Set: 0
Variables
First Row Contains Headers
Variables In Input Data
Selected Variables
Age_08_04
HP
Quarterly_Tax
Weight
CC
Guarantee_Period
PC1
PC2
Weight Variable:
Output Variable:
Price
Help
Cancel
< Back
Next >
Finish
Data
Data Source $A$1:$T$979
Selected Variables Price, Age_08
Partitioning Method Randomly Chosen
Random Seed 12345
Variables 20
Training Rows 587
Validation Rows 391
Test Rows 0
Selected Variables
Price Age_08_04 HP Quarterly_Tax Weight
16900 27 90 210 124
19600 25 192 100 118
22000 28 192 100 118
16950 30 110 85 110
15950 30 110 85 110
15750 29 110 85 112
12950 29 97 19 110
15950 27 97 85 110
PC2 PC3 PC4 PC5 PC6 PC7
PC5 PC6 PC7 PC8 PC9 PC10
0.240055 0.214602 -0.10674 0.269413 -0.7814 -0.4123
1.231269 0.130015 0.212647 1.134958 1.681827 -1.2505
0.784989 1.56114 -0.44845 0.845149 0.579216 -0.8920
-0.22436 -0.86687 0.342644 0.536035 -0.08366 -0.341
0.767042 -0.54424 0.365323 -1.76279 1.163502 -0.3877
-1.83351 0.8969 1.009176 0.485208 0.701223 -1.8458
-0.40576 1.166515 2.173592 1.530927 0.587301 0.5841
-1.70576 0.188072 0.865407 0.127787 -0.69166 -0.8931
Data_Partition
MLR_Output1-Backward
Ready
Search Windows
5:05 PM
3/7/2017

Forward:

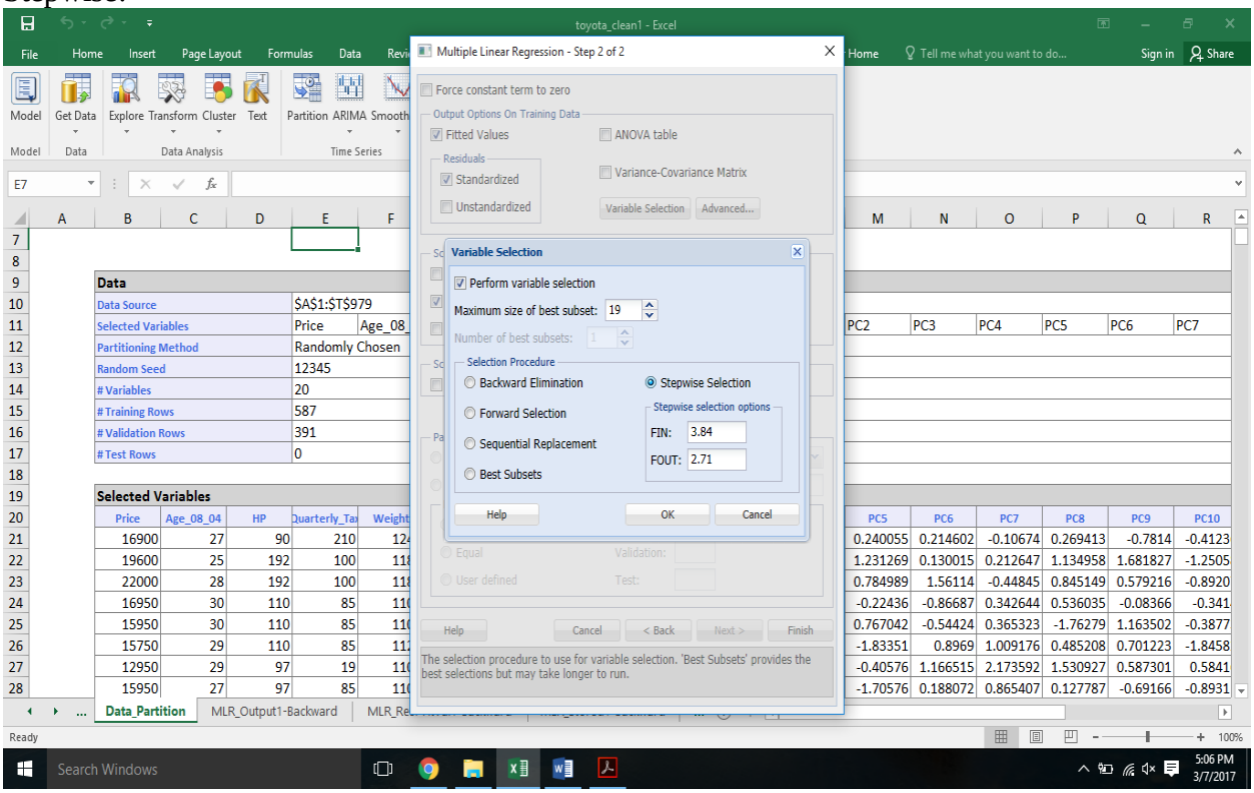

toyota_clean1 - Excel
Multiple Linear Regression - Step 2 of 2
Force constant term to zero
Output Options On Training Data
Fitted Values
ANOVA table
Residuals
Standardized
Unstandardized
Variance-Covariance Matrix
Variable Selection
Advanced...
Variable Selection
Perform variable selection
Maximum size of best subset: 19
Number of best subsets: 1
Selection Procedure
Backward Elimination
Stepwise Selection
Forward Selection
Stepwise selection options
Sequential Replacement
FIN: 3.84
FOUT: 2.71
Best Subsets
Help
OK
Cancel
Equal
Validation:
User defined
Test:
Help
Cancel
< Back
Next >
Finish
The selection procedure to use for variable selection. 'Best Subsets' provides the best selections but may take longer to run.
Data
Data Source $A$1:$T$979
Selected Variables Price, Age_08
Partitioning Method Randomly Chosen
Random Seed 12345
Variables 20
Training Rows 587
Validation Rows 391
Test Rows 0
Selected Variables
Price Age_08_04 HP Quarterly_Tax Weight
16900 27 90 210 124
19600 25 192 100 118
22000 28 192 100 118
16950 30 110 85 110
15950 30 110 85 110
15750 29 110 85 112
12950 29 97 19 110
15950 27 97 85 110
PC2 PC3 PC4 PC5 PC6 PC7
PC5 PC6 PC7 PC8 PC9 PC10
0.240055 0.214602 -0.10674 0.269413 -0.7814 -0.4123
1.231269 0.130015 0.212647 1.134958 1.681827 -1.2505
0.784989 1.56114 -0.44845 0.845149 0.579216 -0.8920
-0.22436 -0.86687 0.342644 0.536035 -0.08366 -0.341
0.767042 -0.54424 0.365323 -1.76279 1.163502 -0.3877
-1.83351 0.8969 1.009176 0.485208 0.701223 -1.8458
-0.40576 1.166515 2.173592 1.530927 0.587301 0.5841
-1.70576 0.188072 0.865407 0.127787 -0.69166 -0.8931
Data_Partition
MLR_Output1-Backward
Ready
Search Windows
5:05 PM
3/7/2017

Archit Jajoo
ISDS 574

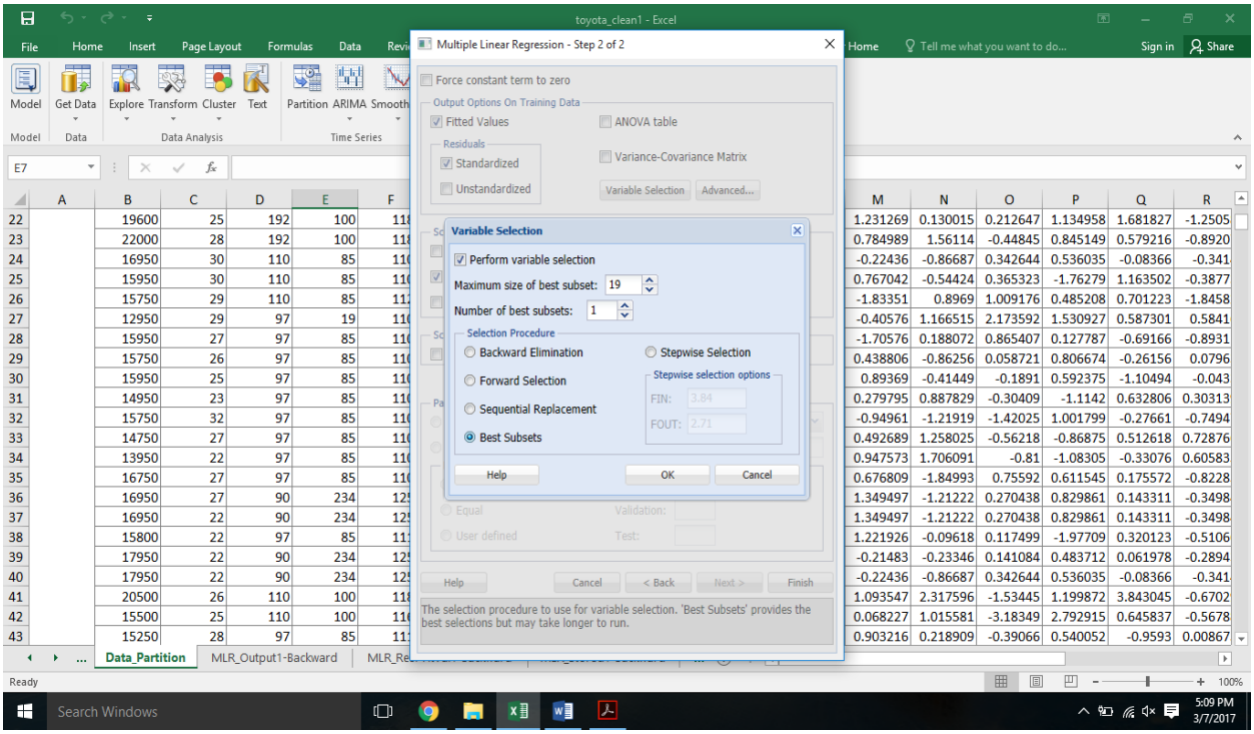
Backward:



Stepwise:



Best Subset:



f)

Forward Selection:

Price = 10429.42 - 151.04* Age_08_04 + 27.08 *HP + 5.43* Weight - 313.12 *PC1 - 131.19 * PC3 + 181.1 *PC4 + 180.64 *PC8 + 386.7 *PC9 -333.12*PC10 + 115.70*PC11

Backward Elimination:

Price = 9898.99 - 145.87* Age_08_04 + 34.5 *HP + 7.00*Quarterly_Tax + 6.5* Weight - 1.58*CC + 60.36*Guarantee_Period - 332.91* PC1 - 52.23*PC2 - 121.5*PC3 + 178.8*PC4 - 54.9*PC5 + 91.6*PC6 - 22.3*PC7 + 176.9*PC8 + 382.28*PC9 - 309.51*PC10 + 129*PC11 + 57.16*PC12 -26.94*PC13

Stepwise Selection:

Price = 10429.42 - 151.04* Age_08_04 + 27.08 *HP + 5.43* Weight - 313.12 *PC1 - 131.19 * PC3 + 181.1 *PC4 + 180.64 *PC8 + 386.7 *PC9 -333.12*PC10 + 115.70*PC11

Best Subset:

Price = 10022.89 - 141.44* Age_08_04 + 36.27 *HP + 8.56*Quarterly_Tax + 6.34* Weight - 1.87*CC - 345.45* PC1 + 204.61*PC4 + 88.98*PC6 + 198.14*PC8 + 366.3*PC9 - 331.077*PC10 + 136.58*PC11 -30.4*PC13

g)
RMSE in validation data

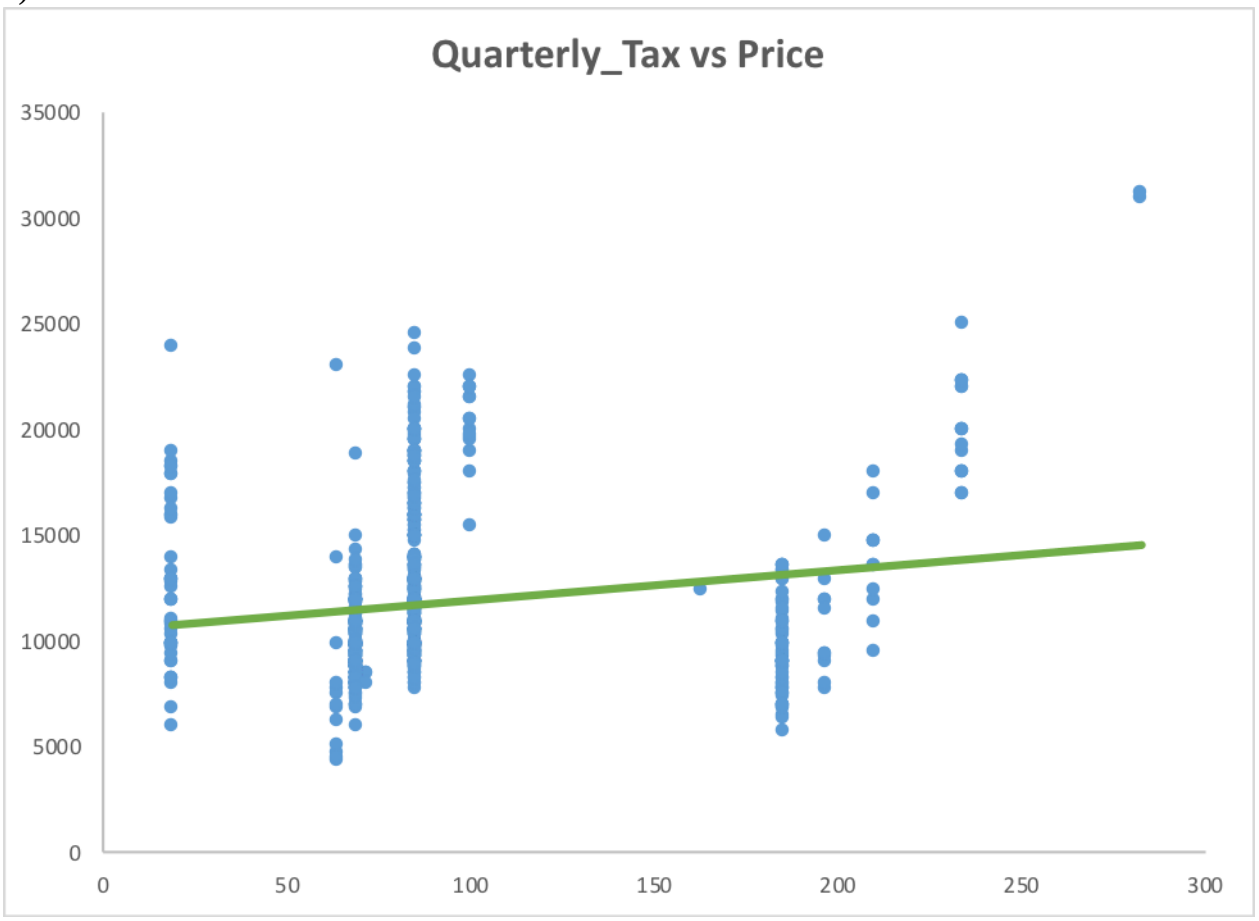
Forward Selection: 1513.01

Backward Elimination: 1408.788

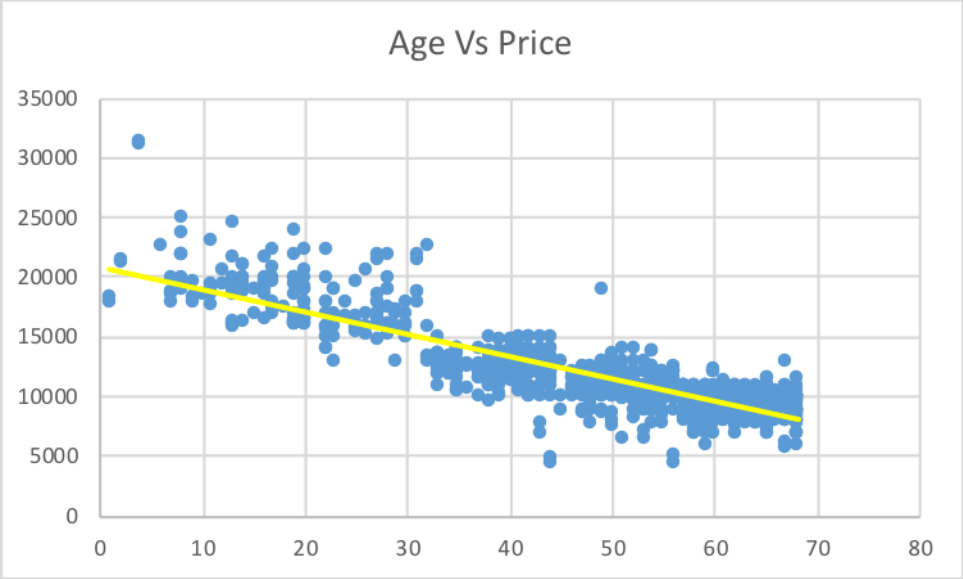
Stepwise Selection: 1513.009

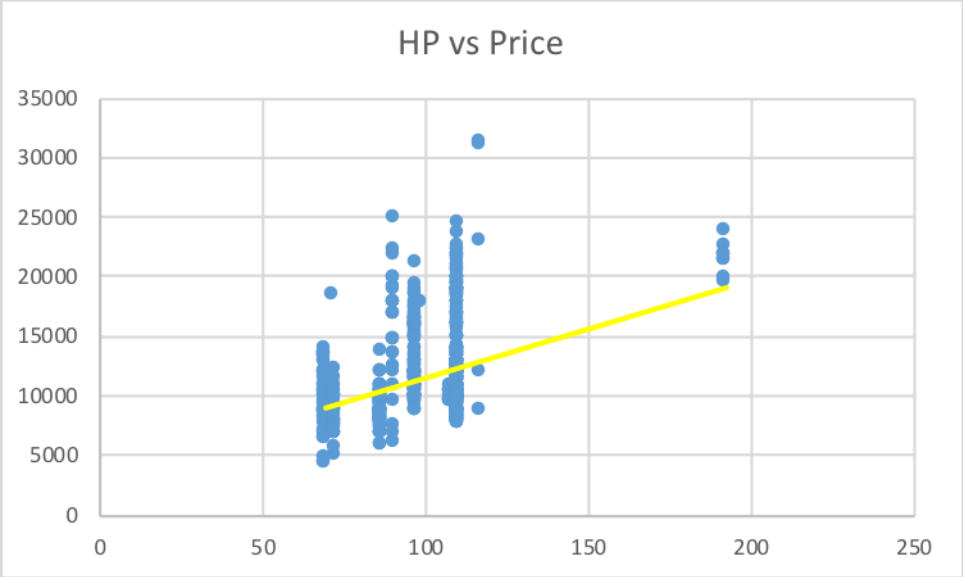
Best Subset: 1405.47

h)



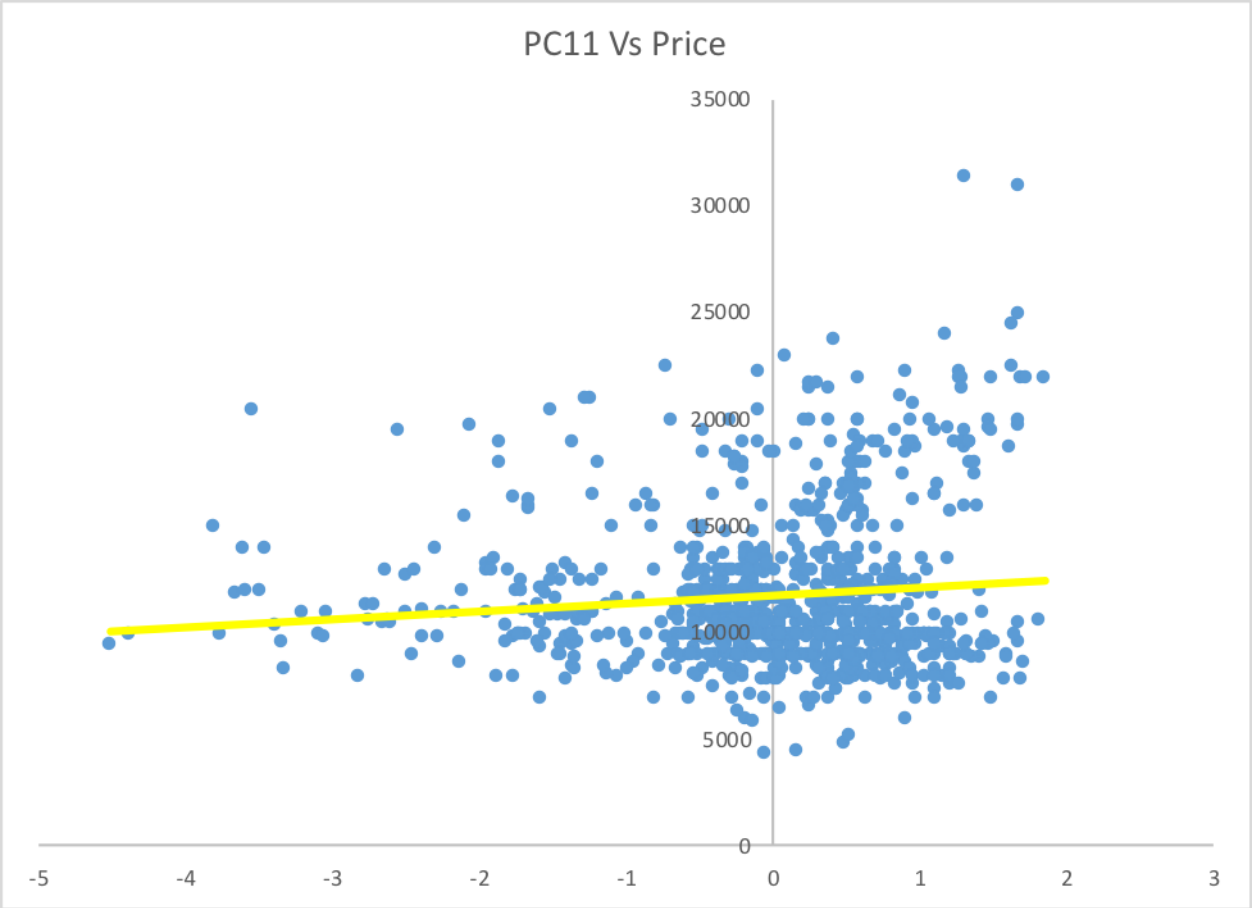
Archit Jajoo
ISDS 574



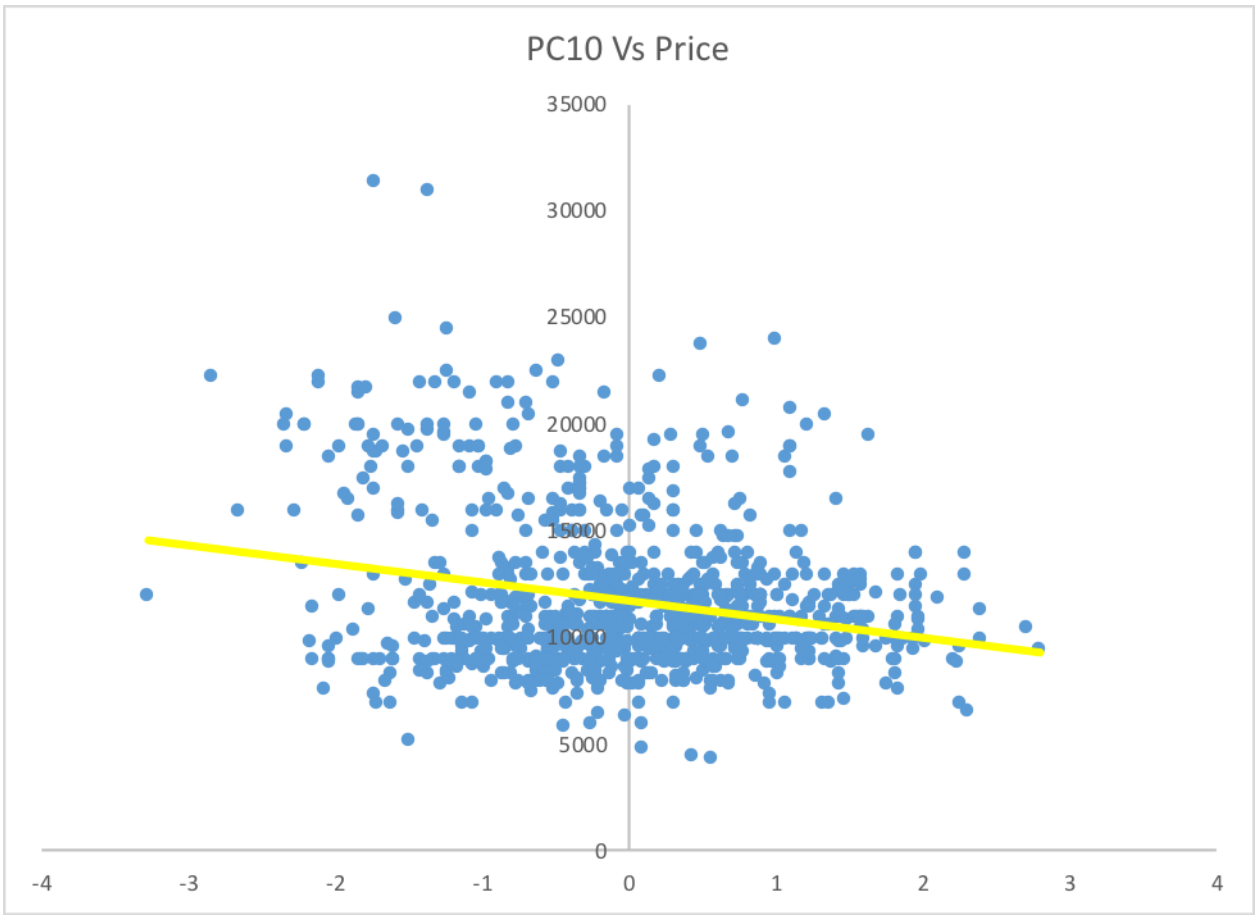


Archit Jajoo
ISDS 574

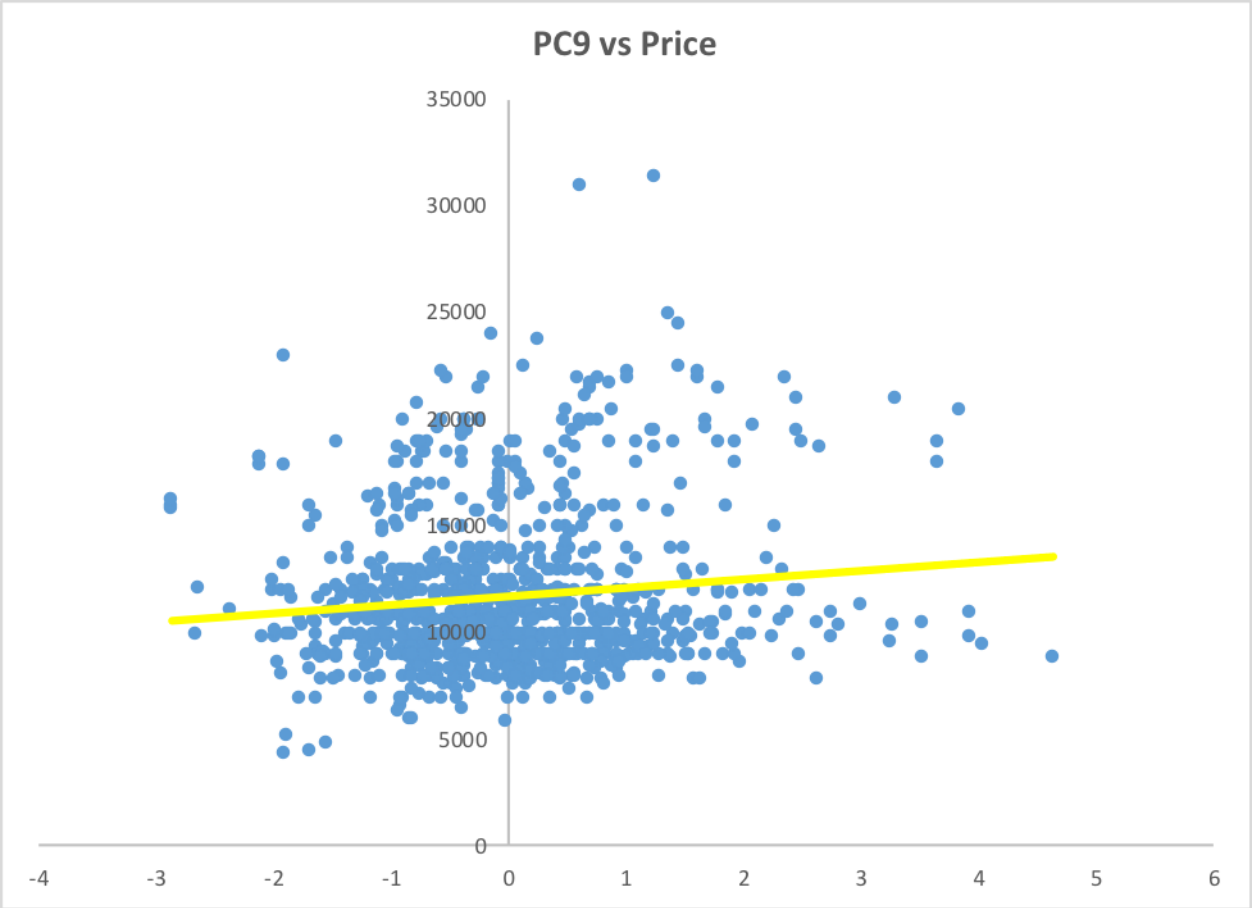
Archit Jajoo
ISDS 574



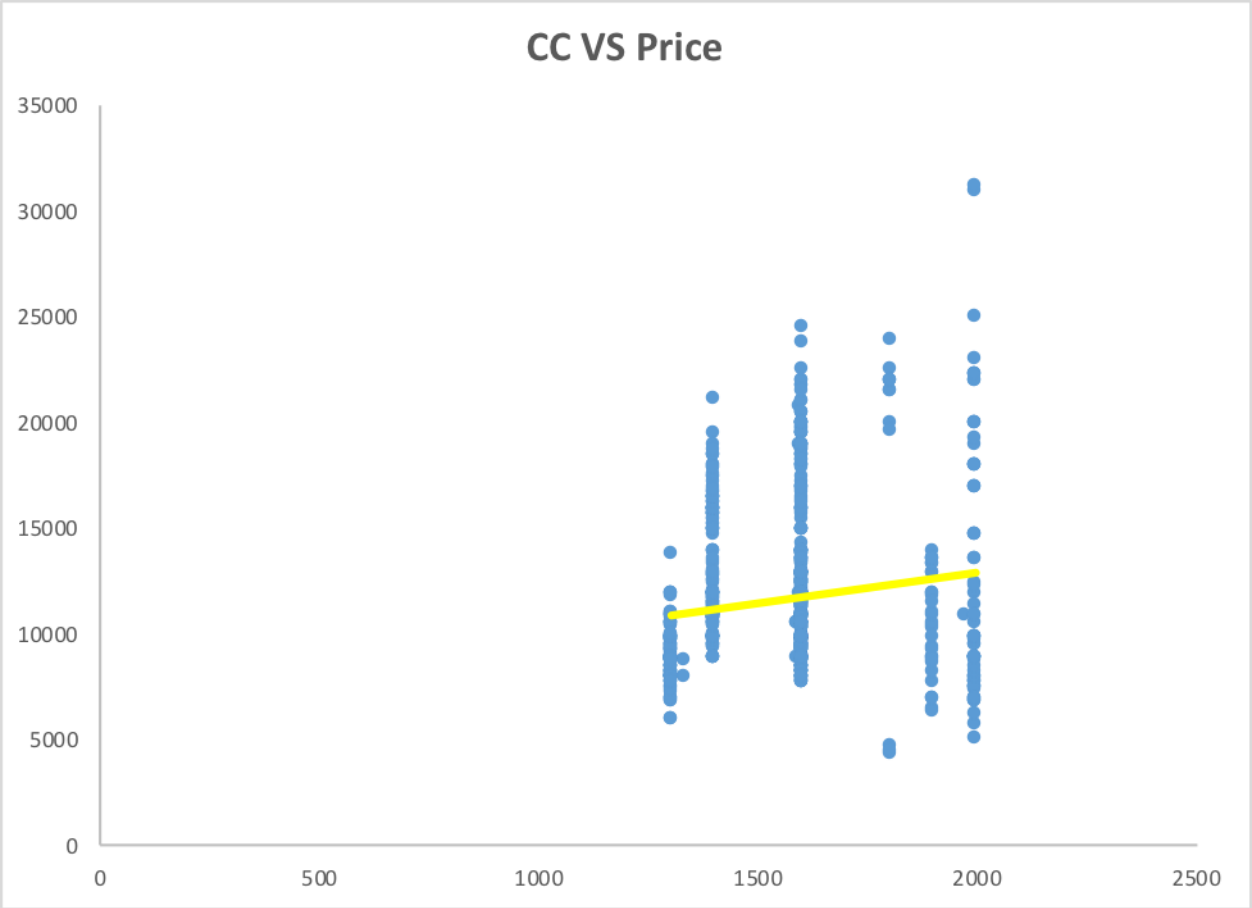
Archit Jajoo
ISDS 574



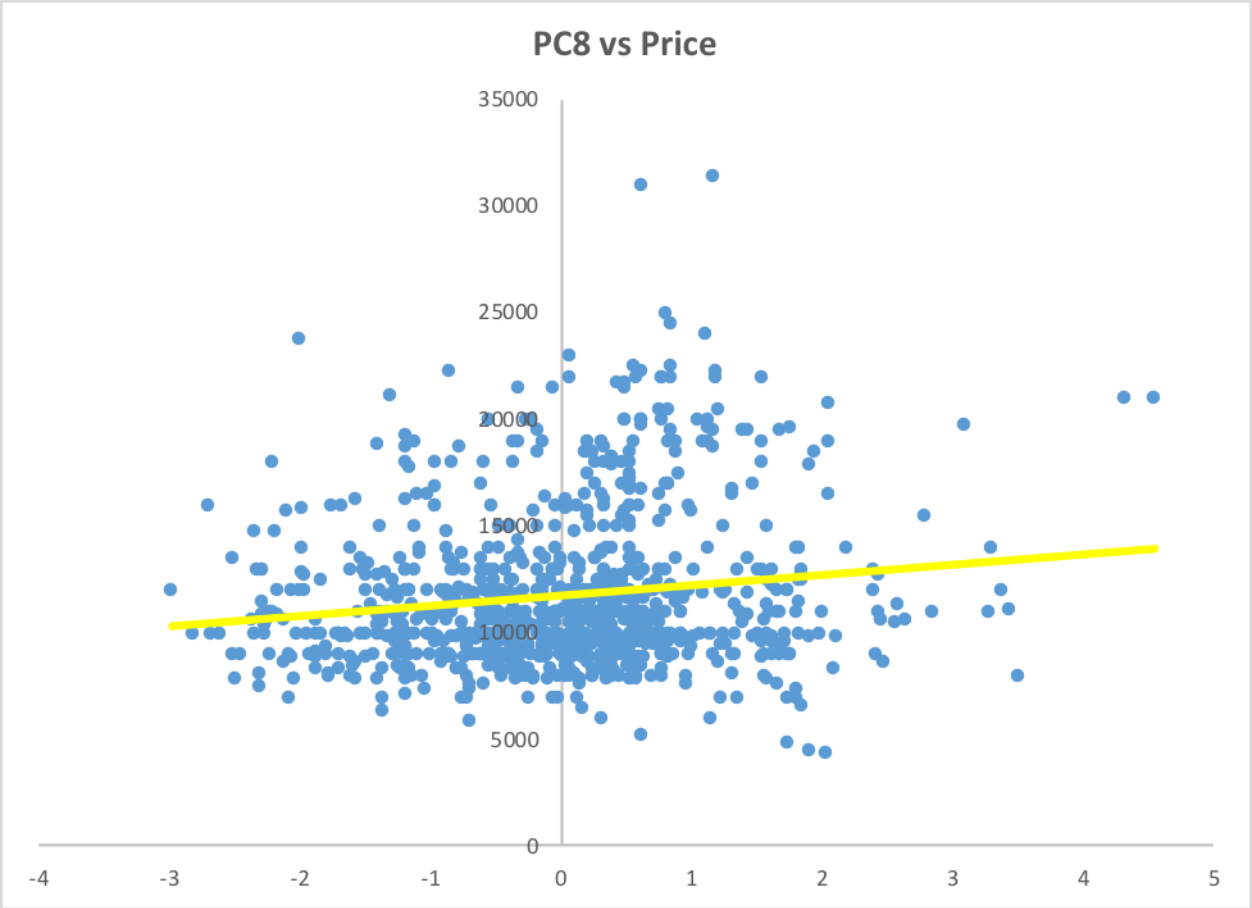
Archit Jajoo
ISDS 574



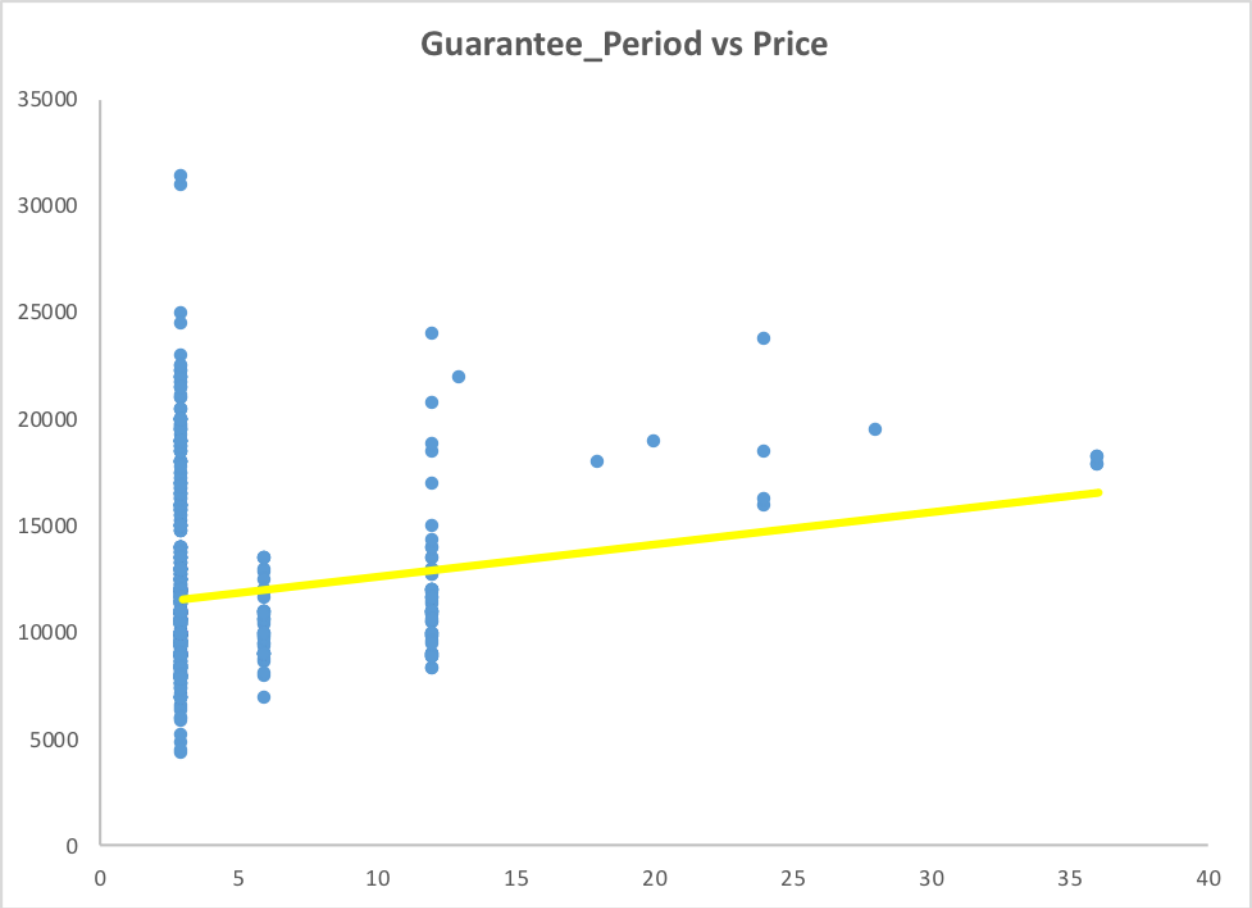
Archit Jajoo
ISDS 574



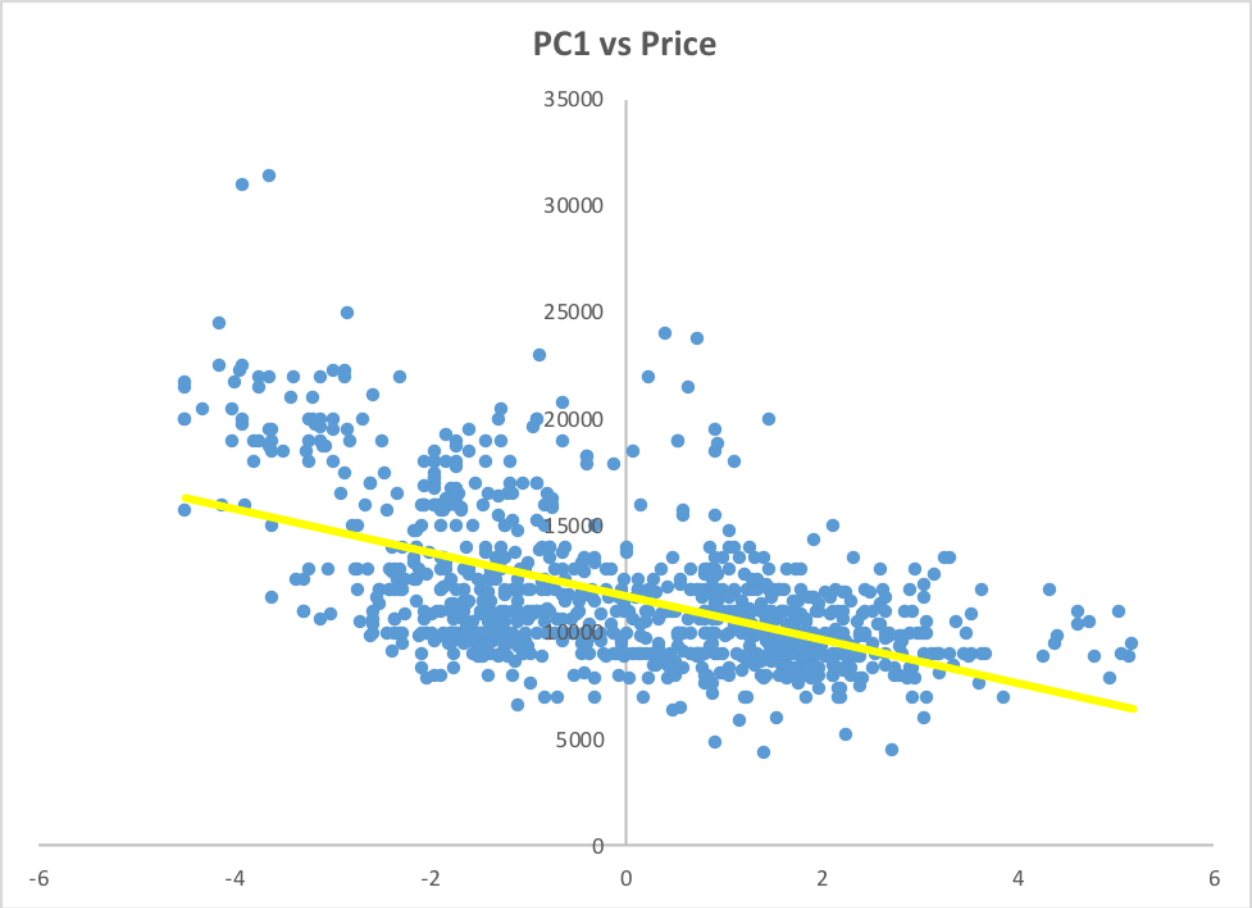
Archit Jajoo
ISDS 574



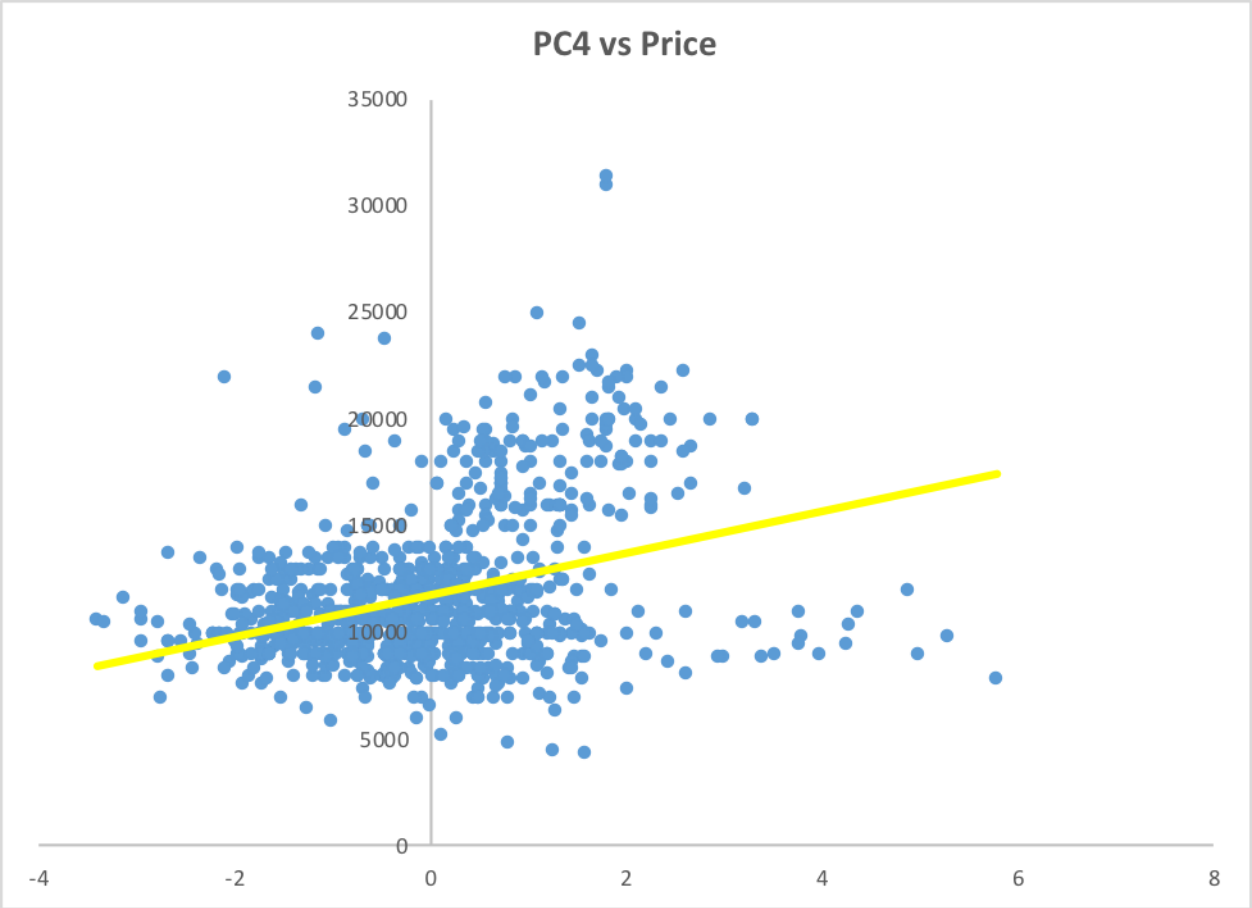
Archit Jajoo
ISDS 574

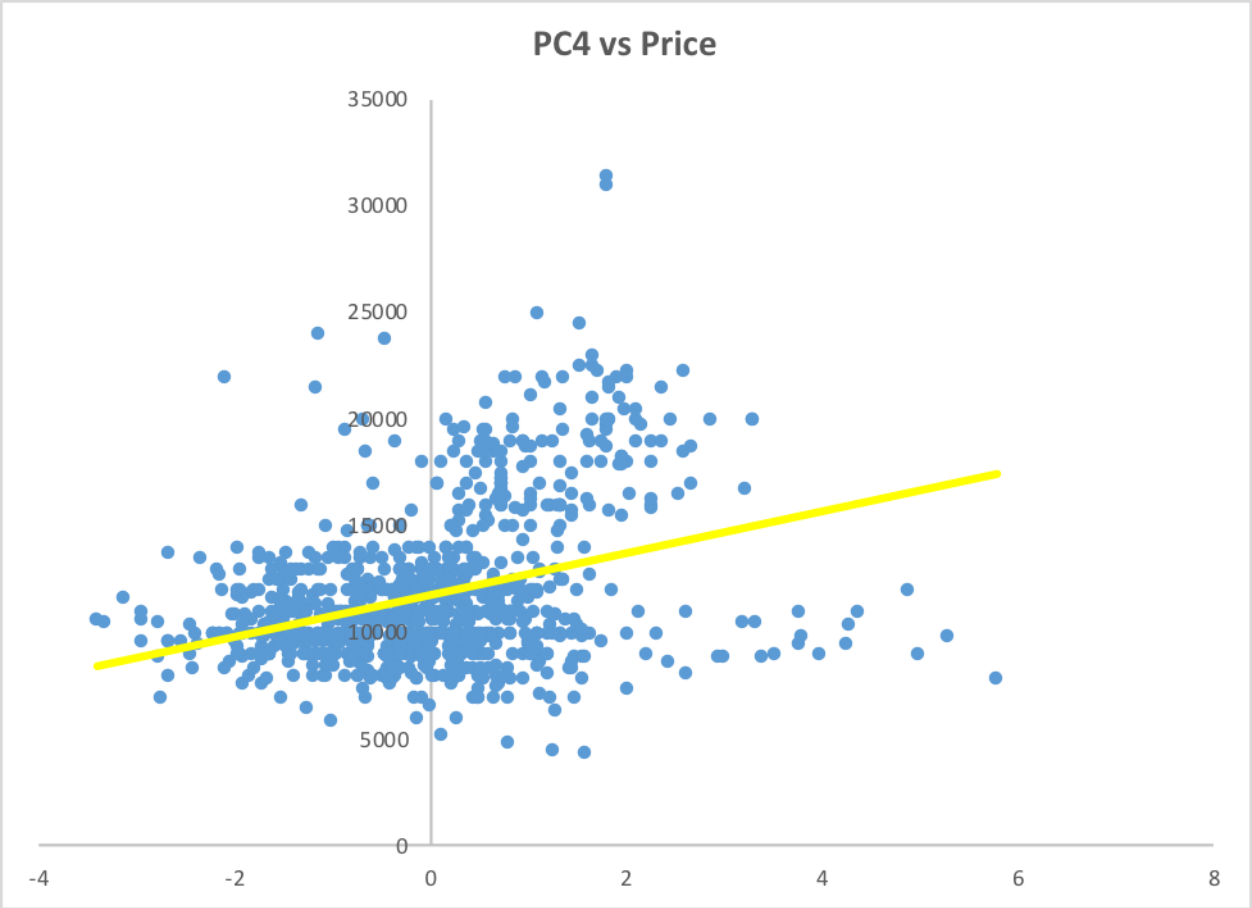


Archit Jajoo
ISDS 574



Archit Jajoo
ISDS 574





i) The assumptions are not satisfied in some cases and in some cases they are. For cases where they are not satisfied we should choose a different subset for all four models and try running them again.