Introduction

Our goal within this project is to analyze the residential building dataset about the real state single-family apartment in Iran. There is a wide range of variables in the dataset such as cost, start year, area, completion year and so on. To analyze the dataset, we use principal component analysis for finding the potential redundancy within the data. Obtained information from PCA would help us to reach better regression models for predicting the outputs.

Data Description and Processing

In the first step, normality of the variables is assessed through the shapiro.test and qqplot. The p-values on normality test for all the variables are less than alpha level equal to 0.05, and it would not be considered as normal. As discussed in the course slides, PCA assumes approximate normality of the input space distribution. Moving the data toward the normal distribution would lead to getting more information. Therefore, log transformation has been used to normalize the data. This transformation normalized a couple of the variables but not all of them. Applying more specific transformation without more details about the distribution of variables and their scales is not a wise decision. Figure 1 shows the normalization on a total floor area of the building variable. During this phase, the output variables are separated to make the data ready for the PCA investigation.

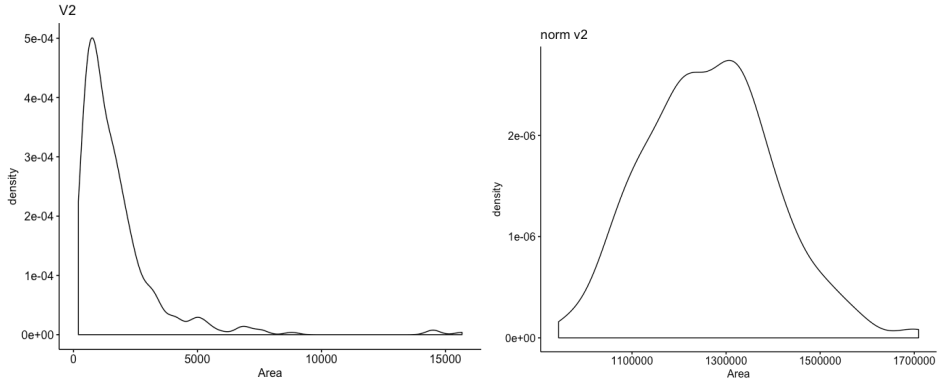

Figure 1

PCA Analysis

PCA analysis could be done using both covariance and correlation matrices. The main reasons that the correlation matrix has been selected for principal component analysis are different scales of variables and highly correlation pattern within the data as shown in Figure 2.

Component	1	2	3	4	5	6	7	8	9	10	11
St. Dev	8.59	2.53	2.16	2.11	1.88	1.72	1.31	1.29	1.13	1.08	0.96
Prop. Var	0.69	0.06	0.043	0.04	0.03	0.03	0.016	0.015	0.015	0.011	0.008
Cum. Var	0.69	0.75	0.79	0.83	0.87	0.89	0.91	0.93	0.94	0.95	0.96

Table 1

To decide the number of PCs, we need to account for how much information regarding the variance would be acquired by selecting each component. As it is shown in Table 1, the first component contributes around 69% of the variance. By increasing the number of components, the cumulative variance would also increase. One possible solution would be to consider all the above eleven components to reach 96% variance of the data. However, this is not the optimal case. We need to consider that from component seven to eleven they are contributing to less than 10% of the variance.

On the other side, increasing the number of components would lead to more coefficients in the regression problem and make the linear model more complicated. The complex model could lead to overfitting problems. According to figure 2, many of the variables are lying toward the same components. The biplot shows scores and loadings of the variables on the first two principal components. As we utilized the correlation matrix the standard deviation of the vectors is equal to 1. Aside from a summary of principal components summarized in Table 1, scree plot of first ten components shows the trend of variance. This plot confirms that by considering the first seven components, we should be able to form a stable baseline to build the regression model on top of that. The arrows in the left upper part of the Figure 2 could be considered as outliers. The loadings show that the PCs we picked can maximize the information we need through the data. As the number of the components are high, the loadings of the seven first components we picked were shown in Appendix, Table 4. By considering all the information gathered from Table 4 and Figure 2, the first seven components would be considered for further PCA based regression analysis.

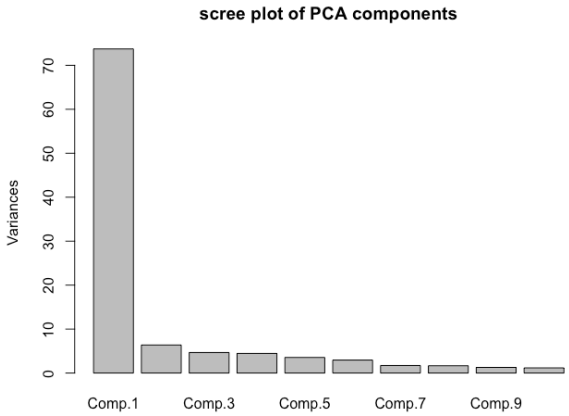

Figure 3

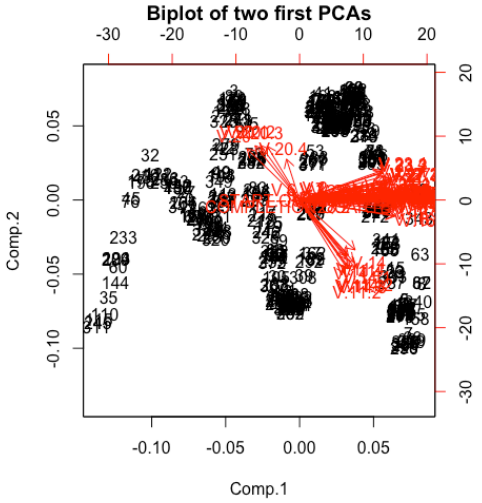

Figure 2

PC based Regression Analysis

Linear regression models have been implemented to predict the actual sales price and actual construction costs of the residential buildings. As discussed in the previous section, seven principal components could help to maximize the information gathered from data. Therefore, all the components would be used in the regression analysis. However, it was taken to an account if it leads to overfitting or not. If it had led to overfitting, the smaller number of components would have been selected to make the model less complicated. The result for linear regression model for predicting the sales price is shown in Table 2. Adjusted R square could help us in understanding if our model fits the data well or not. The goodness of fit we get is 0.597. Or roughly 60% of the variance found in the response variable (sales price) can be explained by the predictor variables. Using more principal components in defining linear models might help in increasing the R-Squared value. However, it could lead to overfitting also. As we do not have holdout test or validation data to assess if overfitting is happening or not, the number of principal components will not change. The actual and predicted results of the linear model plotted in Figure 4 gives a better sense of the model's performance.

Table 2: R-Squared and P-values of linear regression

R-Squared(output1)	P-value(output1)	R-squared(output 2)	P-value(output2)
0.597	<2.2e-16	0.75	<2.2e-16

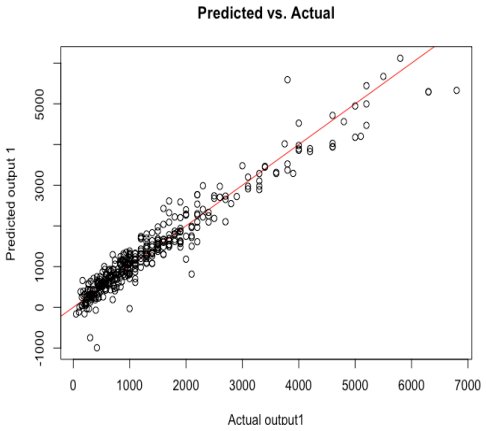

Figure 4

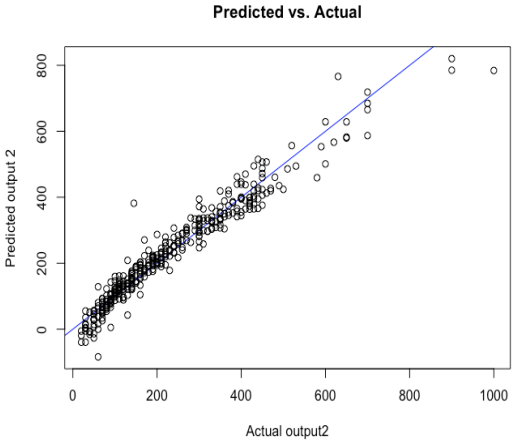

Figure 5

The similar procedure has been used to implement the linear regression model for predicting the actual cost of construction. The statistics we obtained by using first seven components are gathered in Table 3. By considering the same reasons discussed above, we do not decide to incorporate more components in the linear regression process. The plot containing the actual vs predicted results of the actual construction cost is shown in Figure 5. By considering the R^2 in Table 2, 75% of the variance found in the response variable (construction price) can be explained by the predictor variables. This confirms that our linear model performs reasonably well by provided information. The p-values for both linear models give us enough evidence to reject the null hypothesis which shows that there is some relation in the underlying samples which help us in modeling the variables based on the other ones.
P-values of Coefficients of linear regression models are summarized in Table 3. The values lower than 0.05 alpha level shows that component have significant effect on the response variable where as large values suggests that there is no relation between the component and the response variable. According to Table 3, components 1,4,5,6, and 7 have significant effects on the response variable for predicting both actual price and construction cost.

Table 3: Coefficients p-values for output variables

Component name	Coefficients p-values for output 1	Coefficients p-valuesfor output 2
Component 1	<2.2e-16	<2.2e-16
Component 2	0.69764	0.4597
Component 3	0.37753	0.5367
Component 4	0.00127	<2.2e-16
Component 5	<2.2e-16	<2.2e-16
Component 6	1.55e-10	5.49e-10
Component 7	0.05618	0.00543

Appendix

Table 4: PCA loadings for the variables across the first seven components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
START.YEAR	0.11572182	0.01743102	-0.0024579	-0.0170006	-0.0084997	0.03161366
START.QUARTER	-0.0092236	-0.0024339	0.31391347	0.15133566	0.14748061	-0.3146144
COMPLETION.YEAR	0.11483457	0.02192261	0.00658437	-0.0041661	0.0093082	0.00185726
COMPLETION.QUARTER	0.004377	-0.023622	-0.040344	-0.0656829	-0.0659011	0.13387173
V.1	0.00658289	-0.0059371	0.05425809	0.01457606	0.36224053	0.23715341
V.2	0.01456697	0.04008975	-0.0315616	-0.0152175	-0.3780452	-0.2494822
V.3	0.01219846	0.03232798	-0.02408	-0.024837	-0.3759203	-0.242084
V.4	0.07348901	0.04563938	-0.0294408	-0.0107666	-0.3422349	-0.2079196
V.5	0.10026541	0.03209461	-0.0151623	-0.0019726	-0.1654098	-0.0812401
V.6	-0.014959	0.03146274	-0.0775052	0.04826113	-0.3444593	-0.1492468
V.7	0.01198049	0.03547655	-0.0641805	0.03146575	0.04153546	0.03830118
V.8	0.09332762	0.0195436	-0.0264514	0.02654923	-0.1915961	-0.1039507
V.11	0.05616656	-0.2270181	-0.0962222	0.17112135	-0.0160679	0.04992228
V.12	0.11567168	0.01262231	0.02413711	-0.0339336	0.0095522	-0.0056369
V.13	0.11584817	0.01422513	0.01554548	0.00662278	0.0063083	0.00250191
V.14	0.0657792	-0.200409	-0.1126104	0.11985946	-0.0149734	0.02525174
V.15	0.11570736	0.01196531	0.01561088	-0.0307343	0.00709359	0.00432844
V.16	0.11265683	-0.0654096	0.02330774	0.03313917	0.00828021	-0.0161641
V.17	0.11501383	-0.0291096	0.02856531	-0.0467699	0.00218581	0.00526209
V.18	0.06659321	0.03546884	-0.3101392	-0.0454252	0.13761495	-0.1379389
V.19	0.10711083	0.03352793	-0.1315635	-0.0193281	0.06581219	-0.070555
V.20	-0.0623864	0.20637272	-0.1192028	0.2086431	0.00037067	0.01597559
V.21	0.11439935	0.01481183	0.01830133	-0.0258679	0.00981868	0.00070235
V.22	0.11431776	0.00512291	0.02443399	-0.0477675	0.01286379	-0.0095255
V.23	0.09866485	0.06536198	0.07319527	-0.0855826	0.00456575	0.03949223
V.24	0.10058195	0.06582636	-0.0528922	0.17205218	-0.0059022	0.03280195
V.25	0.11593571	0.02346077	0.00061997	0.02188292	0.00068678	0.01367362
V.26	0.11602353	0.01906941	0.00285502	0.01118407	0.0042787	0.00869789
V.27	0.10820654	0.08405237	0.00166337	0.03682743	0.01236567	0.02752126
V.28	0.03334614	0.02805481	-0.3639228	-0.1266374	0.16660408	-0.1956311
V.29	0.11263667	0.02176712	0.01712233	-0.0622125	0.00022213	0.00658527
V.11.1	0.05677527	-0.276562	-0.008824	0.15526579	-0.0093136	-0.0043386
V.12.1	0.11577888	0.01438414	0.01595877	-0.0313266	0.01000431	-0.0063913

V.13.1	0.11578558	0.01858186	-0.0006119	0.01289039	0.01107458	-0.006376
V.14.1	0.06391738	-0.2489834	-0.0042017	0.14635188	-0.0230558	0.00608123
V.15.1	0.11571925	0.01011071	0.02470137	-0.02937	0.00438249	0.00637892
V.16.1	0.11268192	-0.0626382	-0.0152597	0.00561294	0.01704112	-0.0139287
V.17.1	0.11522031	-0.019123	0.02380699	-0.0464215	0.00552527	-0.0029375
V.18.1	0.06643931	0.02088345	-0.0748966	-0.0329403	-0.165806	0.30813774
V.19.1	0.10748357	0.01094137	-0.0276208	-0.0162194	-0.0670969	0.12975417
V.20.1	-0.0527643	0.2185878	-0.0972082	0.26326206	-0.0207192	0.07444775
V.21.1	0.1143849	0.01560418	0.01430177	-0.0265805	0.00713478	0.0035476
V.22.1	0.11436505	0.00492425	0.01614408	-0.0478806	0.00817416	-0.003713
V.23.1	0.09852308	0.09470467	0.06457877	-0.1246392	0.02726312	0.00393168
V.24.1	0.10120376	0.05533233	-0.0592047	0.17957431	-0.000363	0.028168
V.25.1	0.11584731	0.02482337	0.00271467	0.03016336	0.00445229	0.00859081
V.26.1	0.11589315	0.02124488	0.00510349	0.01268402	0.0016823	0.01325107
V.27.1	0.10870847	0.0856592	-0.0046259	0.02909483	0.01781178	0.0186529
V.28.1	0.02588359	-9.55E-05	-0.0806706	-0.1150138	-0.2210534	0.36725912
V.29.1	0.11300221	0.02119002	0.03298181	-0.0326955	0.00775893	-0.0020903
V.11.2	0.05427926	-0.2951919	-0.0763325	0.10901946	0.0328249	-0.0453264
V.12.2	0.11581334	0.01517278	0.01521177	-0.0262002	0.00554448	0.0002967
V.13.2	0.11549203	0.02184589	-0.0106545	0.02165077	0.00415686	0.00318996
V.14.2	0.06559708	-0.2720693	-0.0745428	0.0855657	0.02806255	-0.052911
V.15.2	0.11582327	0.00965606	0.02126623	-0.0296658	0.00774265	0.00170513
V.16.2	0.11261755	-0.0403386	-0.0014694	0.00457932	0.00073274	0.01598722
V.17.2	0.11524982	-0.00917	0.02296493	-0.0438663	0.00396711	-0.0046631
V.18.2	0.06892246	0.00869227	0.1323886	0.13619117	-0.0786957	0.16812432
V.19.2	0.10879288	-0.0036274	0.06034481	0.05589618	-0.0324353	0.07019035
V.20.2	-0.0461481	0.2200137	-0.0799506	0.29907909	-0.0096547	0.04754606
V.21.2	0.11414282	0.01420772	0.01571259	-0.0140869	0.00593406	0.00196199
V.22.2	0.1145524	0.01534641	0.02905085	-0.036285	0.00981791	-0.0064169
V.23.2	0.09837832	0.10954656	0.06184039	-0.1445358	0.02024399	0.01314239
V.24.2	0.10136443	0.04537361	-0.0654543	0.18525604	0.00560699	0.02633246
V.25.2	0.11573909	0.02686204	-0.0029376	0.0339351	0.00872029	0.00421614
V.26.2	0.1157123	0.02467673	0.00789814	0.01583897	0.00213714	0.01206198
V.27.2	0.10910125	0.08353394	2.34E-05	0.01778821	0.01424077	0.02078843
V.28.2	0.02521865	-0.0204993	0.1965521	0.11270485	-0.1236449	0.19347445
V.29.2	0.11308722	0.01903412	0.01840462	-0.0187967	0.02275578	-0.0228185
V.11.3	0.05780457	-0.2751037	-0.0355981	0.06316847	-0.0406788	0.02888204
V.12.3	0.11581814	0.01687513	0.01689063	-0.016406	0.00399473	0.00083305
V.13.3	0.11500262	0.02645494	-0.0131892	0.03661667	0.00050689	0.00932244

V.14.3	0.05963034	-0.2773618	-0.018688	0.08031573	-0.0108189	-0.0043728
V.15.3	0.11584482	0.00971228	0.02294689	-0.0299963	0.00707852	0.00095499
V.16.3	0.11231114	-0.0242494	0.00849672	0.01301949	-0.0087129	0.0241297
V.17.3	0.1152452	0.00331391	0.01759384	-0.0419135	0.003014	-0.0025478
V.18.3	0.06278478	0.01612478	0.23956769	0.17354209	0.11864734	-0.2162355
V.19.3	0.10669054	-0.0075844	0.11498332	0.07483193	0.04786235	-0.0944753
V.20.3	-0.0383449	0.2161012	-0.0807411	0.31684315	0.00856152	0.01248673
V.21.3	0.11407286	0.01245904	0.02193353	-0.0158284	-0.0035215	-0.0070345
V.22.3	0.11439417	0.01516077	0.0183923	-0.0228827	0.0109753	-0.0167844
V.23.3	0.09841453	0.1127458	0.06324208	-0.1467744	0.01557722	0.0188571
V.24.3	0.10186038	0.03386281	-0.0749524	0.18585871	-0.0078683	0.03071586
V.25.3	0.11564612	0.02748277	-0.0106304	0.03515773	0.00399646	0.01071634
V.26.3	0.11557044	0.02699647	0.00681336	0.01714263	0.00273307	0.00968676
V.27.3	0.10943033	0.08211397	0.00758268	0.01211195	0.00960025	0.01670772
V.28.3	0.01584785	0.0010869	0.31464677	0.14935081	0.14625455	-0.3091337
V.29.3	0.11308596	0.01031102	-0.0071249	-0.0207707	0.01441898	-0.0121291
V.11.4	0.05759543	-0.2213758	-0.0426916	0.05098498	-0.0448461	0.03179782
V.12.4	0.11576697	0.01775753	0.01713225	-0.0089568	0.00608727	-0.0034265
V.13.4	0.11419701	0.03050516	-0.0094806	0.05454829	0.00383668	0.00045422
V.14.4	0.05495736	-0.235202	-0.0816779	0.04139882	-0.0267368	0.01065374
V.15.4	0.11591416	0.01078769	0.01554679	-0.0317558	0.00635671	0.00422108
V.16.4	0.11130616	-0.0037743	0.02444796	0.01576104	0.0079829	-0.0079335
V.17.4	0.11522173	0.01867028	0.02252056	-0.0329904	0.00285728	-0.0001931
V.18.4	0.07346233	0.02929135	-0.2918402	-0.043892	0.1345766	-0.1226909
V.19.4	0.10886737	-0.0114237	-0.123056	-0.0260098	0.06511596	-0.0617272
V.20.4	-0.0155432	0.16423559	-0.1564989	0.3122443	0.02611556	0.03416547
V.21.4	0.11366832	0.00812476	0.01422748	-0.0159741	0.01154653	-0.0012819
V.22.4	0.11448817	0.02592044	0.00661101	-0.0210492	0.00981483	-0.0098881
V.23.4	0.09796303	0.11865241	0.07874258	-0.1438101	0.0131576	0.00884814
V.24.4	0.10258946	0.02233288	-0.0606033	0.18965989	-0.0096587	0.03619504
V.25.4	0.11556385	0.02479125	-0.0052482	0.03938752	-9.33E-05	0.01686292
V.26.4	0.11552946	0.0268805	0.00372811	0.01483005	0.00196187	0.01180804
V.27.4	0.10966165	0.07612992	0.02046451	0.00236634	0.00960195	0.01117007
V.28.4	0.03363712	0.02852479	-0.3634798	-0.1267374	0.16648079	-0.1954956
V.29.4	0.11346105	-0.0004014	-0.0002477	-0.0064776	-0.0021436	0.00683942