MTECH KE5107

DATA MINING METHODOLODY AND METHODS

PROJECT REPORT

PRINCIPAL COMPONENT ANALYSIS

CLUSTER ANALYSIS

REGRESSION ANALYSIS

TEAM MEMBERS
SIDDHARTH PANDEY
PRANSHU RANJAN SINGH
EDUARD ANTHONY CHAI
NYON YAN ZHENG
TAN KOK KENG

MASTER OF TECHNOLOGY IN
KNOWLEDGE ENGINEERING
BATCH KE-30(2018)

1.0 DATA UNDERSTANDING

1.1 DATA COLLECTION

We used the data provided from one of Kaggle Competition, Sberbank Russian Housing Market, which can be downloaded here. The aim of this competition is to predict the sale price of property in Russia. Data was collected by Sberbank, Russia's oldest and largest bank. The data consists of property transactions in Moscow, Russia from August 2011 to June 2015.

1.2 DATA EXPLORATION

The dataset consists of 292 variables and 30,471 observations. From 292 variables, 276 variables are continuous variable and 16 variables are categorical. (See Appendix for Data Description)

1.3 DATA PREPARATION

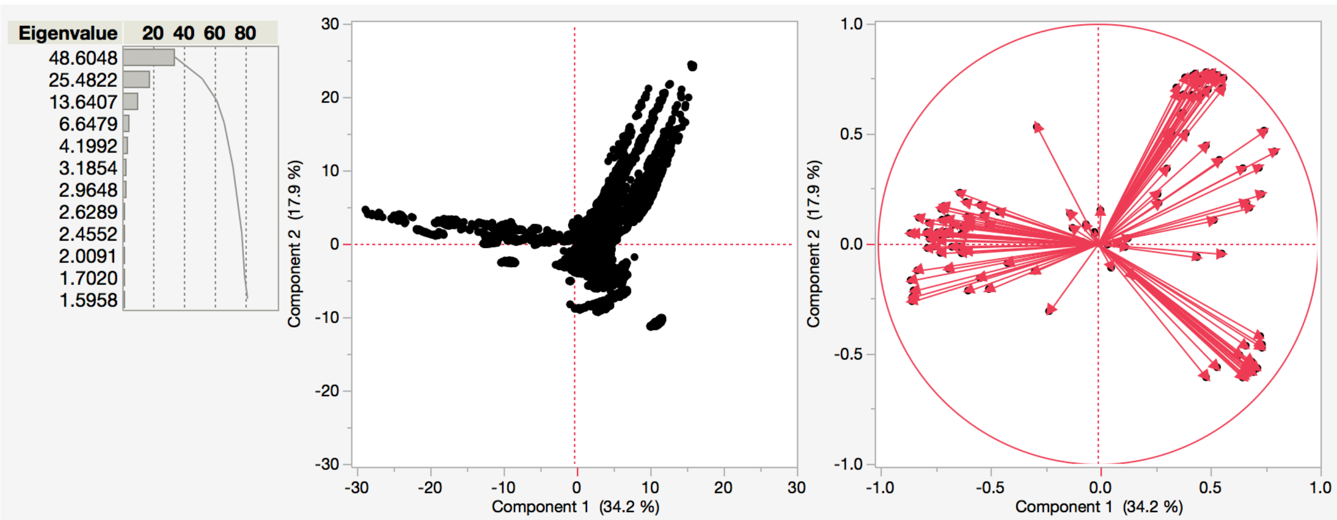
To prepare the data for analysis, we started by removing categorical variables and variables that only consists IDs. Then we removed variables that have more than 15% missing values. We also removed observations that have more than 30% missing values. The rest of the data with missing values were imputed using different techniques such as mean, median and linear model. Then, we looked for outliers and removed them. Going through this process left us with 28,942 observations and 153 variables.

2.0 SOFTWARE & TOOLS

- Data exploration and preparation: R Studio
- Statistical software: JMP Pro 13

3.0 PRINCIPAL COMPONENTS ANALYSIS

3.1 PRINCIPAL COMPONENTS EXTRACTION



We performed the components extraction after considering the 4 criteria below:

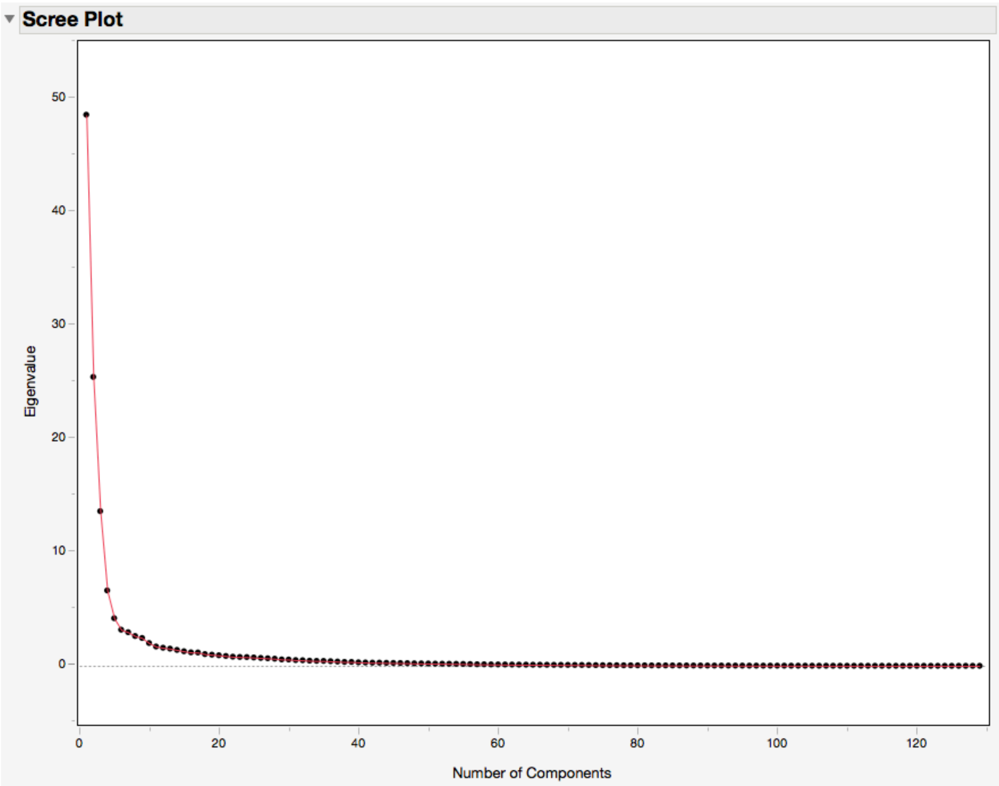
1. **Eigenvalues > 1**
   Using the eigenvalue criterion, we should be extracting 18 components. However, this figure is still quite high considering the components 11 to 18 only explain less than 2% of variance each.

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	48.6048	34.229	34.229
2	25.4822	17.945	52.174
3	13.6407	9.606	61.780
4	6.6479	4.682	66.462
5	4.1992	2.957	69.419
6	3.1854	2.243	71.662
7	2.9648	2.088	73.750
8	2.6289	1.851	75.601
9	2.4552	1.729	77.330
10	2.0091	1.415	78.745
11	1.7020	1.199	79.944
12	1.5958	1.124	81.068
13	1.5150	1.067	82.135
14	1.4022	0.987	83.122
15	1.2847	0.905	84.027
16	1.1896	0.838	84.864
17	1.1753	0.828	85.692
18	1.0343	0.728	86.420
19	0.9815	0.691	87.112

2. **The percentage of variance criterion**
   We could see that we needed 12 components to explain at least 80% of the variance and at least 8 components to explain 75% of the variance.

3. **Scree plot**
   The scree plot showed that the first "elbow" occurred between the 9th and 10th component.



4. **Interpretability of the components**
   We looked into the rotated loading matrix and we are able to interpret the component until component 10th. More than 10, we find that is either the components are insignificant (low absolute loadings values) or it is hard to interpret.

After weighting all the considerations above, we decided to extract 10 components for our analysis and this would explain 78.745% of the variance.

3.2 PRINCIPAL COMPONENTS INTERPRETATION

We considered correlation above 0.4 in absolute values to help us interpret the components.

Principal Component 1: Population

Features	Factor 1
young_male	0.978278378
young_all	0.978134238
X0_13_male	0.978052558
X0_17_all	0.978020617
X0_17_male	0.977898834
X0_13_all	0.977830362
X0_17_female	0.976239517
young_female	0.975942576
X0_6_male	0.975729213
X0_6_all	0.975490771
children_preschool	0.975490771
X0_13_female	0.975399383
X7_14_male	0.975045572
children_school	0.974183029
X7_14_all	0.974183029
X0_6_female	0.973428124
X7_14_female	0.969926735
work_female	0.944829706
work_all	0.942246436
raion_popul	0.941579154
work_male	0.934544373
school_quota	0.906779566
preschool_quota	0.89207823
ekder_female	0.848520938
preschool_education_centers_raion	0.83756121
ekder_all	0.836733116
school_education_centers_raion	0.806433266
ekder_male	0.800891396
shopping_centers_raion	0.525336918
sport_objects_raion	0.499659101
additional_education_raion	0.479532164
healthcare_centers_raion	0.437759691
market_count_5000	0.236716986

Principal component 1 explained **34.229%** of the variance. The highly loaded values were referring to the population of young, working adult and the elderly. This also explained the reason why some of the essentials amenities for these groups of the population were also loaded in this component, such as education centres were for the younger population and health care was for the elderly population. We named this component **Population**.

Principal Component 2: Number of amenities

Features	Factor 2
church_count_3000	0.96529521
big_church_count_3000	0.96237084
church_count_2000	0.956605785
big_church_count_2000	0.955813495
church_count_1500	0.948825318
big_church_count_1500	0.94657327
office_raion	0.945246015
church_count_1000	0.933016305
big_church_count_1000	0.928703379
leisure_count_5000	0.928117091
cafe_count_5000	0.926675206
church_count_5000	0.923866855
office_count_5000	0.922865445
big_church_count_5000	0.92281848
cafe_count_5000_price_1000	0.91655925
office_sqm_3000	0.886322386
church_count_500	0.884459455
big_church_count_500	0.87825812
culture_objects_top_25_raion	0.862687232
office_sqm_5000	0.838352544
office_sqm_2000	0.835540979
office_sqm_1000	0.801818836
office_sqm_1500	0.799211226
sport_count_5000	0.724152022
university_top_20_raion	0.723574896
trc_sqm_1000	0.708349658
trc_sqm_500	0.696009142
trc_count_5000	0.681284912
trc_sqm_1500	0.679043159
sport_objects_raion	0.653191879
shopping_centers_raion	0.635205205
trc_sqm_2000	0.624248708
mosque_count_3000	0.601665711
trc_sqm_3000	0.598101339
office_sqm_500	0.595142433
trc_sqm_5000	0.571097952
additional_education_raion	0.505051825
mosque_count_5000	0.49649531
mosque_count_2000	0.466665288
mkad_km	0.429825087
bulvar_ring_km	-0.433901566
kremlin_km	-0.457055496

Principal component 2 explained **17.945%** of the variance. The highly loaded values were referring to the number of amenities. It also had a negative correlation with the distance to kremlin. It makes sense because we should find more amenities when we are closer to the city. We named this component **Number of amenities**.

Principal Component 3: Distance from metro and city

Features	Factor 3
power_transmission_line_km	0.911566707
radiation_km	0.904603217
metro_km_walk	0.889322423
metro_min_walk	0.889322423
metro_km_avto	0.884834613
park_km	0.877540554
metro_min_avto	0.853130501
ts_km	0.828243904
thermal_power_plant_km	0.810417996
mosque_km	0.801099342
exhibition_km	0.800410351
ttk_km	0.755725759
basketball_km	0.752940682
stadium_km	0.739203712

sadovoe_km	0.72205457
incineration_km	0.717316392
bus_terminal_avto_km	0.71364407
kremlin_km	0.708105059
bulvar_ring_km	0.706456888
big_market_km	0.701845405
oil_chemistry_km	0.691443683
mkad_km	0.674551822
zd_vokzaly_avto_km	0.661851438
nuclear_reactor_km	0.603598626
big_road2_km	0.541749452
detention_facility_km	0.530369065
big_church_km	0.49247858
workplaces_km	0.408865597
swim_pool_km	0.392116953

Principal component 3 explained **9.606%** of the variance. This component refers to the distance from some of the energy facilities, metro, city, and some amenities. We named this component **Distance from metro and city**.

Principal Component 4: Distance from railroad and public facilities

Features	Factor 4
railroad_station_avto_km	0.88975208
railroad_station_walk_min	0.884671068
railroad_station_walk_km	0.884671068
railroad_km	0.852870042
railroad_station_avto_min	0.831094402
school_km	0.827989537
preschool_km	0.826768405
shopping_centers_km	0.762488627
public_healthcare_km	0.730056182
big_church_km	0.693850639
swim_pool_km	0.681152139
area_m	0.644203343
public_transport_station_min_walk	0.629457082

public_transport_station_km	0.629457082
museum_km	0.56196822
office_km	0.544767585
workplaces_km	0.54211114
ice_rink_km	0.539322115
kindergarten_km	0.535680665
detention_facility_km	0.49925101
green_part_5000	0.470003545
university_km	0.448081821
fitness_km	0.435862913
additional_education_km	0.434173409
basketball_km	0.425771561
market_shop_km	0.412913549
theater_km	0.402577757
green_zone_part	0.395580125

Principal component 4 explained **4.682%** of the variance. Most of the highly loaded values were explaining about the distance from railroad stations and public facilities. We named this component **Distance from railroad station and public facilities**.

Principal Component 5: Distance from cultural amenities

Features	Factor 5
university_km	0.699198921
theater_km	0.686965058
museum_km	0.614489214
hospice_morgue_km	0.461636083
workplaces_km	0.4368465
sport_count_5000	-0.403152106
prom_part_5000	-0.44639639
market_count_5000	-0.55517825

Principal component 5 explained **2.957%** of the variance. Most of the highly loaded values were referring to the distance from the university, theater and museum. We named this component **Distance from cultural amenities**.

Principal Component 6: Distance from water treatment facilities

Features	Factor 6
water_treatment_km	0.775059732
cemetery_km	-0.425481943
big_road1_km	-0.43062894

Principal component 6 explained **2.243%** of the variance. In this component, distance from water treatment facilities was prominent. It also referred to the distance to big road. We named this component **Distance from water treatment facilities.**

Principal Component 7: Number of mosques

Features	Factor 7
mosque_count_1500	0.816408512
mosque_count_1000	0.763509922
mosque_count_2000	0.668161638
mosque_count_3000	0.516829635
mosque_count_500	0.516648668
mosque_count_5000	0.437652223
school_education_centers_top_20_raion	0.242405671

Principal component 7 explained **2.088%** of the variance. All of the loaded features were referring to the number of mosques. We named this component **Number of Mosques.**

Principal Component 8: Greenery

Features	Factor 8
green_zone_part	0.632642411
green_part_5000	0.547374582
indust_part	-0.612138484

Principal component 8 explained **1.851%** of the variance. Most of the highly loaded values were referring to the proportion of greenery area. This also explained the negative correlation with industry area. We named this component **Greenery.**

Principal Component 9: Size of property

Features	Factor 9
full_sq	0.905330936
life_sq	0.864238196
num_room	0.829182514
kitch_sq	0.477873605
max_floor	0.38524752
floor	0.357662141

Principal component 9 explained **1.729%** of the variance. Most of the highly loaded values were referring to the characteristics of the house such as total area, living area, number of rooms, kitchen area and etc. We named this component **Size of property.**

Principal Component 10: Shopping mall density

Features	Factor 10
trc_sqm_1000	0.414930506
trc_sqm_500	0.386565546
trc_sqm_1500	0.37164371
church_count_500	0.27437056

Principal component 10 explained **1.415%** of the variance. Only one feature was loaded here which was referring to the shopping mall area within 1,000 metres. We named this component **Shopping mall density**.

4.0 CLUSTER ANALYSIS

4.1 VALIDATION STRATEGY

We used the Development and Validation as our validation strategy. Hence, the data was split into 50:50.

4.2 CLUSTERING TECHNIQUE

We used the k-means clustering technique in this analysis. We set the clustering algorithm to have maximum of 7 clusters.

4.3 CLUSTERS INTERPRETATION

The cluster analysis was performed first on the development dataset.

The table below showed the Cubic Clustering Criterion (CCC) scores for the various clustering solutions. Larger values of CCC indicate better fit.

Method	Cluster	CCC	Best
K-Means Clustering	3	9.67	
K-Means Clustering	4	21.49	
K-Means Clustering	5	53.28	
K-Means Clustering	6	128.62	
K-Means Clustering	7	186.46	Optimal CCC

K-MEANS 6-CLUSTER SOLUTION

In this analysis, we limited our acceptable clustering solution to a maximum of 6 clusters. Starting with the 6-cluster solution which had the best CCC fit, we took a quick look at the number of observations and cluster means in the table below. Each of the 6 clusters count had substantial numbers of observations which did not indicate any outliers forming a cluster by itself.

Cluster	Number of Observations	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
1	540	4.11	-0.8	1.42	1	-1.32	5.52	1.64	-3.75	-0.21	-0.2
2	268	16.11	24.07	10.6	0.63	-4.81	-2.63	-1.04	0.2	-1.32	1.63
3	537	-18.25	2.88	7.32	4.14	3.61	0.78	-1.36	0.72	0.69	0.1
4	1329	6.35	5.72	-1.66	-0.9	2.86	0.79	0.22	0.74	1.25	-1.37
5	7663	-2.21	-0.35	-0.95	-0.21	-0.34	-0.14	-0.11	0.15	-0.97	-0.15
6	4134	2.72	-2.87	0.35	-0.03	-0.21	-0.66	0.19	-0.14	1.38	0.63

The highlighted cluster 2 which was a small cluster did not look coherent relative to the descriptions of Principal 1 (high population) and Principal 3 (distance to city centre and metro). It showed that cluster 2 observations were highly populated but far from the city centre and metro. We expected municipalities with high population to be closer to the city centre and transportation.

Therefore, **we decided to rule out this solution** and proceed to look at the solution with the next best CCC fit which is the 5-cluster solution.

K-MEANS 5-CLUSTER SOLUTION

The table below showed the cluster count and means for each cluster for the development dataset. Each of the 5 clusters count had substantial numbers of observations which did not indicate any outliers forming a cluster by itself.

Cluster	Number of Observations	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
1	516	4.1	-0.86	1.5	0.98	-1.36	5.6	1.65	-3.91	-0.15	-0.28
2	537	-18.25	2.88	7.32	4.14	3.61	0.78	-1.36	0.72	0.69	0.1
3	7894	-1.61	0.46	-0.54	-0.17	-0.51	-0.21	-0.14	0.15	-1	-0.08
4	4164	2.71	-2.85	0.33	-0.04	-0.21	-0.65	0.18	-0.14	1.37	0.62
5	1360	6.3	5.66	-1.68	-0.89	2.83	0.78	0.21	0.74	1.22	-1.34

As we can see, the cluster means for each principal component of the development dataset. On the surface, there appeared to be substantial differences between cluster means to proceed with profiling. There was no obvious incoherence between the clusters means for each of the cluster.

Therefore, **we have decided to proceed with the analysis using this solution.**

4.4 CLUSTERS VALIDATION

We compared the cluster solutions from the development and validation dataset for consistency with respect to the:
- cluster count
- cluster mean

The table below showed the cluster count for each of the 5 clusters for the both the development and validation dataset. The distribution of number of observations in each cluster in both the development and validation set appeared similar.

Cluster	Dataset	Count
1	development	516
2	development	537
3	development	7894
4	development	4164
5	development	1360

Cluster	Dataset	Count
1	validation	543
2	validation	560
3	validation	7676
4	validation	4392
5	validation	1300

The table below showed the cluster means for each principal component for the both datasets. The cluster means in each cluster for the development and validation dataset appeared similar.

Cluster	Dataset	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
1	development	4.1	-0.86	1.5	0.98	-1.36	5.6	1.65	-3.91	-0.15	-0.28
2	development	-18.25	2.88	7.32	4.14	3.61	0.78	-1.36	0.72	0.69	0.1
3	development	-1.61	0.46	-0.54	-0.17	-0.51	-0.21	-0.14	0.15	-1	-0.08
4	development	2.71	-2.85	0.33	-0.04	-0.21	-0.65	0.18	-0.14	1.37	0.62
5	development	6.3	5.66	-1.68	-0.89	2.83	0.78	0.21	0.74	1.22	-1.34
1	validation	4.34	-1.14	1.97	0.91	-1.54	5.51	1.37	-4.09	0.07	-0.52
2	validation	-18.18	2.86	7.29	4.26	3.59	0.79	-1.48	0.68	0.67	0.17
3	validation	-1.58	0.54	-0.59	-0.23	-0.52	-0.2	-0.17	0.17	-1	-0.1
4	validation	2.81	-2.96	0.46	-0.01	-0.21	-0.66	0.23	-0.12	1.36	0.59
5	validation	6.26	5.61	-1.67	-0.86	2.73	0.78	0.25	0.83	1.17	-1.23

Based on the comparison above, we were able to conclude that this cluster solution is stable.

4.5 CLUSTERS PROFILING

The table below showed the cluster profiles across the principal components. The cluster means were binned using an equal size binning method to obtain the categorical values (low, quite low, medium, quite high, high).

Components	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Population	4.1	-18.25	-1.61	2.71	6.3
	quite high	low	medium	quite high	high
Number of Amenities	-0.86	2.88	0.46	-2.85	5.66
	quite low	quite high	quite low	low	high
Distance from metro and city	1.5	7.32	-0.54	0.33	-1.68
	quite near	far	near	quite near	near
Distance from railroad and public facilities	0.98	4.14	-0.17	-0.04	-0.89
	medium	far	quite near	quite near	near
Distance from cultural amenities	-1.36	3.61	-0.51	-0.21	2.83
	near	far	quite near	quite near	quite far
Distance from water treatment facilities	5.6	0.78	-0.21	-0.65	0.78
	far	quite near	near	near	quite near
Number of mosques	1.65	-1.36	-0.14	0.18	0.21
	high	low	medium	medium	medium
Greenery	-3.91	0.72	0.15	-0.14	0.74
	low	high	quite high	quite high	high
Size of property	-0.15	0.69	-1	1.37	1.22
	quite small	quite big	small	big	big
Shopping mall density	-0.28	0.1	-0.08	0.62	-1.34
	medium	quite high	medium	high	low

<u>Summary</u>

We summarised the clusters profile using only those features where their pattern or distribution was markedly different with other clusters.

Cluster	Description	Details
1	Property in industrial area with Muslim communities	▪ near to amenities like university, theatre, museum ▪ far from water treatment plant ▪ many mosques ▪ poor greenery, industrial area
2	Countryside property	▪ in a low population municipality ▪ far from city center, metro, railroad station and public transport ▪ far from amenities like university, theatre, museum ▪ few mosques ▪ good greenery
3	City center small property	▪ near city center and metro ▪ near to water treatment plant

		▪ small property size
4	Suburban large property	▪ in a quite high population municipality with few amenities ▪ near to water treatment plant ▪ large property ▪ high shopping mall density
5	City center large property	▪ in a high population municipality with many amenities ▪ near city center, metro, railroad station and public transport ▪ good greenery ▪ large property ▪ low shopping mall density

5.0 REGRESSION MODEL

5.1 REGRESSION WITHOUT PRINCIPAL COMPONENTS

We attempted to fit a linear regression without using the principal components and we obtained the following summary of fit.

Summary of Fit	
RSquare	0.262911
RSquare Adj	0.260484
Root Mean Square Error	0.522016
Mean of Response	15.61218
Observations (or Sum Wgts)	28942

5.2 REGRESSION WITH PRINCIPAL COMPONETNS

5.2.1 Data Transformation

As we will be using the principal components as predictors to run our regression models, we analysed the mean, standard deviation, skewness and kurtosis for each component and the target variable (price_doc) in the table below.

	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**	**Prin9**	**Prin10**	**price_doc**
Mean	~0	~0	~0	~0	~0	~0	~0	~0	~0	~0	7134326.2
Std Dev	6.97	5.05	3.69	2.58	2.05	1.78	1.72	1.62	1.57	1.42	4604420.7
Skewness	-0.7	2.05	1.45	-2.08	-0.26	1.6	0.8	-0.3	0.26	0.05	3.5
Kurtosis	0.61	8.57	1.71	9.27	0.33	5.47	2.17	3.42	-0.15	1.58	24.3

From the table above, there were a few variables that were skewed (skewness > 0.8). The skewness will be corrected by applying a suitable transformation to each variable. As the principal component scores included negative values, the values were translated to a positive range of values by adding a constant to the values prior to applying the transformation. The transformed variables were standardised again after the transformation. The table below showed the distributions after the transformation.

	SQRT of Prin2	**Log of Prin3**	**Square of Prin4**	**Log of Prin6**	**Log of price_doc**
Mean	~0	~0	~0	~0	~0
Std Dev	1	1	1	1	1
Skewness	0.04	-0.53	0.35	-0.01	-0.74
Kurtosis	4.32	5.5	2.68	2	2.19

5.2.2 LINEAR REGRESSION WITH PCA

With the Principal Component Analysis, we reduced the complexity of the model and at the same time saw improvement in R^2 from 0.262911 (model without using the principal components) to 0.295508 below.

Response norm log price

Effect Summary

Source	LogWorth	PValue
Prin7	1432.671	0.00000
Prin1	473.789	0.00000
Prin8	308.342	0.00000
norm log pos prin6	195.731	0.00000
norm log pos prin3	12.380	0.00000
Prin9	8.943	0.00000
norm sqrt pos prin2	7.375	0.00000
norm sq pos prin4	4.265	0.00005
Prin5	2.416	0.00383
Prin10	1.113	0.07715

Remove Add Edit ☐ FDR

Summary of Fit

RSquare	0.295508
RSquare Adj	0.295265
Root Mean Square Error	0.839386
Mean of Response	9.419e-5
Observations (or Sum Wgts)	28940

5.2.3 POLYNOMIAL REGRESSION WITH PCA

We further fit a third degree polynomial regression model and we saw that the R^2 improved from 0.295508 (above) to 0.313059.

Summary of Fit

RSquare	0.313059
RSquare Adj	0.312418
Root Mean Square Error	0.829108
Mean of Response	9.419e-5
Observations (or Sum Wgts)	28940

5.2.4 SUMMARY

Comparing the 2 models that we have fitted, the polynomial model has a better fit in terms of a higher R^2 and a lower root mean square error. However, an R^2 is generally not good enough a fit to be deployed in a production environment.

A more complex technique may be needed to improve the model fitting. Given the number of variables in the dataset, using a more advanced dimensional reduction techniques might help, such as T-SNE. Also, a more complex model may also improve the prediction further, such as support vector machine and neural networks. More variables can also be added to help improving the model further, such as data from economy and financial sector.

REFERENCES

Market, S. R. (n.d.). *Sberbank Russian Housing Market*. Retrieved February 20, 2018, from Kaggle: https://www.kaggle.com/c/sberbank-russian-housing-market

APPENDIX: DATA DESCRIPTION

Features	Description
0_13_all	Population aged 0-13
0_13_female	Female population aged 0-13
0_13_male	Male population aged 0-13
0_17_all	Population aged 0-17
0_17_female	Female population aged 0-17
0_17_male	Male population aged 0-17
0_6_all	Population aged 0-6
0_6_female	Female population aged 0-8
0_6_male	Male population aged 0-7
7_14_all	Population aged 7-14
7_14_female	Female population aged 7-14
7_14_male	Male population aged 7-14
additional_education_km	Distance to additional education
additional_education_km	Distance to additional education
area_m	Area mun. area, sq.m.
basketball_km	Distance to the basketball courts
big_church_count_1000	The number of big churches in 1000 metres zone
big_church_count_1500	The number of big churches in 1500 metres zone
big_church_count_2000	The number of big churches in 2000 metres zone
big_church_count_3000	The number of big churches in 3000 metres zone
big_church_count_500	The number of big churches in 500 metres zone
big_church_count_5000	The number of big churches in 5000 metres zone
big_church_km	Distance to large church
big_market_km	Distance to grocery / wholesale markets
big_road1_km	Distance to Nearest major road
big_road2_km	The distance to next distant major road
bulvar_ring_km	The distance to the Boulevard Ring
bus_terminal_avto_km	Distance to bus terminal (avto)
cafe_count_5000	The number of cafes or restaurants in 5000 metres zone
cafe_count_5000_price_1000	Cafes and restaurant bill, average 500-1000 in 5000 metres zone
catering_km	Distance to catering
cemetery_km	Distance to the cemetery
children_preschool	Number of pre-school age population
church_count_1000	The number of churches in 1000 metres zone
church_count_1500	The number of churches in 1500 metres zone
church_count_2000	The number of churches in 2000 metres zone
church_count_3000	The number of churches in 3000 metres zone
church_count_500	The number of churches in 500 metres zone
church_count_5000	The number of churches in 5000 metres zone
church_synagogue_km	Distance to Christian churches and Synagogues
culture_objects_top_25_raion	Number of objects of cultural heritage
detention_facility_km	Distance to detention facility
ekder_all	Population older than working age
ekder_female	Female population older than working age

ekder_male	Male population older than working age
exhibition_km	Distance to exhibition
fitness_km	Distance to fitness
floor	for apartments, floor of the building
full_sq	total area in square meters, including loggias, balconies and other non-residential areas
green_part_5000	The share of green zones in 5000 metres zone
green_zone_km	Distance to green zone
green_zone_part	Proportion of area of greenery in the total area
green_zone_part	Proportion of area of greenery in the total area
healthcare_centers_raion	Number of healthcare centres in district
hospice_morgue_km	Distance to hospice/morgue
ice_rink_km	Distance to ice palace
incineration_km	Distance to the incineration
indust_part	Share of industrial zones in area of the total area
industrial_km	Distance to industrial
kindergarten_km	Distance to kindergarten
kitch_sq	kitchen area
kremlin_km	Distance to the city centre (Kremlin)
leisure_count_5000	The number of leisure facilities in 5000 metres zone
life_sq	living area in square meters, excluding loggias, balconies and other non-residential areas
market_count_5000	The number of markets in 5000 metres zone
market_shop_km	Distance to markets and department stores
max_floor	number of floors in the building
metro_km_avto	Distance to subway by car, km
metro_km_walk	Distance to the metro, km
metro_min_avto	Time to subway by car, min.
metro_min_walk	Time to metro by foot
mkad_km	Distance to MKAD (Moscow Circle Auto Road)
mosque_count_1000	The number of mosques in 1000 metres zone
mosque_count_1500	The number of mosques in 1500 metres zone
mosque_count_2000	The number of mosques in 2000 metres zone
mosque_count_3000	The number of mosques in 3000 metres zone
mosque_count_500	The number of mosques in 500 metres zone
mosque_count_5000	The number of mosques in 5000 metres zone
mosque_km	Distance to mosques
museum_km	Distance to museums
nuclear_reactor_raion	Presence of existing nuclear reactors
num_room	number of living rooms
office_count_5000	The number of office space in 5000 metres zone
office_km	Distance to business centres/ offices
office_raion	Number of office space in district
office_sqm_1000	The square of office space in 1000 metres zone
office_sqm_1500	The square of office space in 1500 metres zone
office_sqm_2000	The square of office space in 2000 metres zone
office_sqm_3000	The square of office space in 3000 metres zone
office_sqm_500	The square of office space in 500 metres zone

office_sqm_5000	The square of office space in 5000 metres zone
oil_chemistry_km	Distance to dirty industries
park_km	Distance to park
power_transmission_line_km	Distance to power transmission line
preschool_education_centers_raion	Number of pre-school institutions
preschool_km	Distance to preschool education organizations
preschool_quota	Number of seats in pre-school organizations
price_doc	sale price (this is the target variable)
prom_part_5000	The share of industrial zones in 5000 metres zone
public_healthcare_km	Distance to public healthcare
public_transport_station_km	Distance to the public transport station (walk)
public_transport_station_min_walk	Time to the public transport station (walk)
radiation_raion	Presence of radioactive waste disposal
railroad_km	Distance to the railway/Moscow Central Ring/open areas Underground
railroad_station_avto_km	Distance to the railroad station (avto)
railroad_station_avto_min	Time to the railroad station (avto)
railroad_station_walk_km	Distance to the railroad station (walk)
railroad_station_walk_min	Time to the railroad station (walk)
raion_popul	Number of municipality population. district
sadovoe_km	Distance to the Garden Ring
school_education_centers_raion	Number of high school institutions
school_education_centers_top_20_raion	Number of high schools of the top 20 best schools in Moscow
school_km	Distance to high school
school_quota	Number of high school seats in area
shopping_centers_km	Distance to shopping centres
shopping_centers_raion	Number of malls and shopping centres in district
sport_count_5000	The number of sport facilities in 5000 metres zone
sport_objects_raion	Number of higher education institutions
stadium_km	Distance to stadium
swim_pool_km	Distance to swimming pool
theater_km	Distance to theatre
thermal_power_plant_km	Distance to thermal power plant
trc_count_5000	The number of shopping malls in 5000 metres zone
trc_sqm_1000	The square of shopping malls in 1000 metres zone
trc_sqm_1500	The square of shopping malls in 1500 metres zone
trc_sqm_2000	The square of shopping malls in 2000 metres zone
trc_sqm_3000	The square of shopping malls in 3000 metres zone
trc_sqm_500	The square of shopping malls in 500 metres zone
trc_sqm_5000	The square of shopping malls in 5000 metres zone
ts_km	Distance to power station
ttk_km	Distance to the TTC (Third Transport Ring)
university_km	Distance to universities
university_top_20_raion	Number of higher education institutions in the top ten ranking of the Federal rank
water_km	Distance to the water reservoir / river
water_treatment_km	Distance to water treatment

work_all	Working-age population
work_female	Female working-age population
work_male	Male working-age population
workplaces_km	Distance to workplaces
young_all	Population younger than working age
young_female	Feale population younger than working age
young_male	Male population younger than working age
zd_vokzaly_avto_km	Distance to train station