Predicting House Prices in King County, Seattle

Shraddha Barde
Nidhi Galmale
Amey Thombre

Under the noble guidance of professor Khasha Dehnad

BIA – 652 Multivariate Data Analysis – I

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Hello !

Nidhi Galmale

Amey Thombre

Shraddha Barde

1

Introduction

- Problem Statement
- Objective
- Roadmap
- Potential of the data

INTRODUCTION

Problem Statement

Purchasing a house is a well prevailing necessity or more importantly a dream of every individual.

The real estate market is very dynamic and discrete in nature in terms of pricing of houses.

There always exists a necessity of systems that enables us in predicting acute prices of houses depending on a variety of factors.

INTRODUCTION

Objective

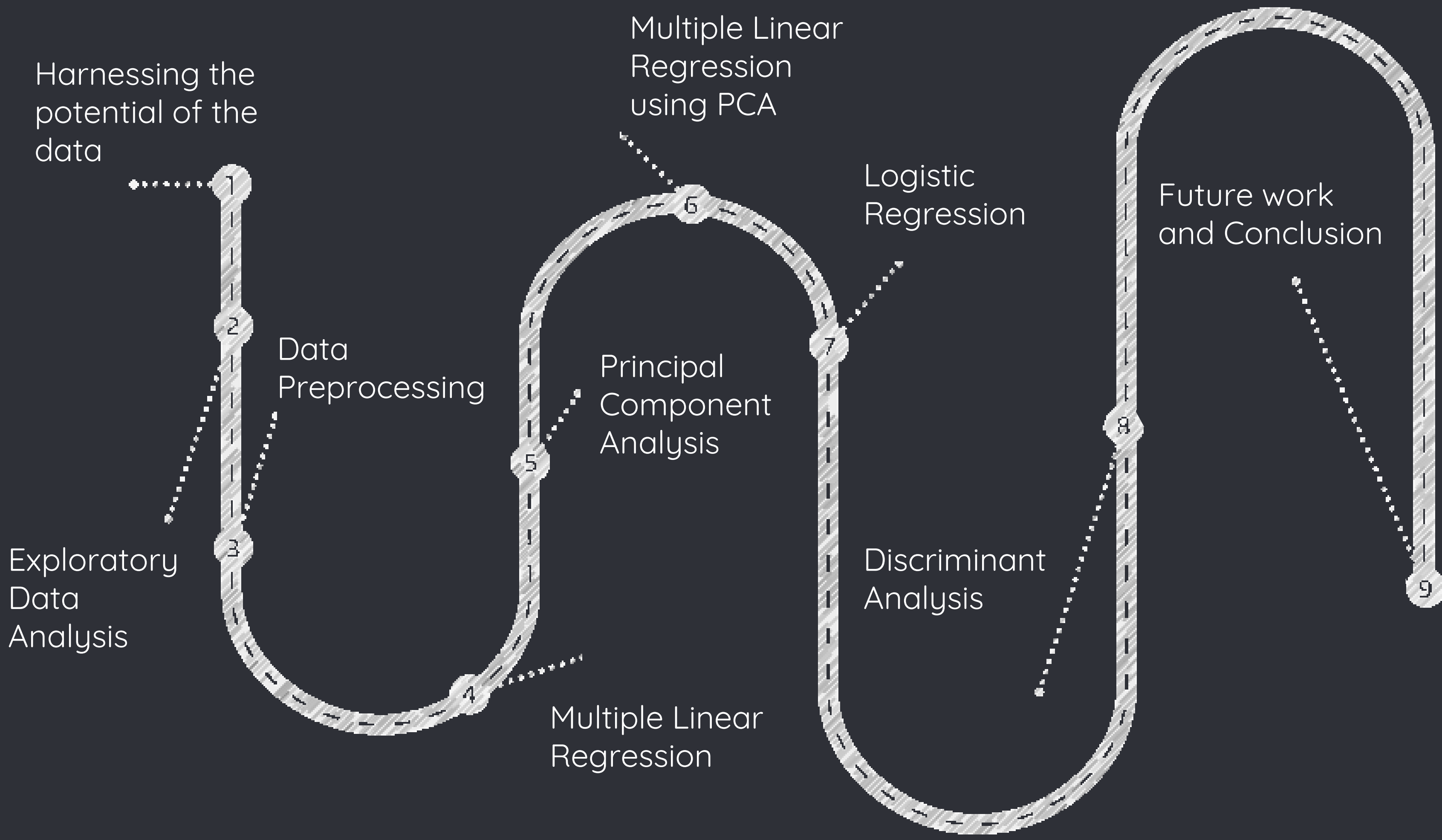
Construct a model that will enable predict prices of houses in a particular area (We have considered the case of King County, Seattle)

This predictive price is a result of a variety of factors such as the view, rooms, area, year built and many other attributes present in the dataset.

Another objective is to determine whether the price of the house what we have deduced is expensive or not expensive.

INTRODUCTION

Roadmap



INTRODUCTION

Potential of the Data

Dataset retrieved from : https://www.kaggle.com/harlfoxem/housesalesprediction

This data set comprised of 19 house features plus the price and the id columns, along with 21613 observations.

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Target variable is price

2

Exploratory Data Analysis

The place to begin !

Exploratory Data Analysis

Exploratory Data Analysis

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
bedrooms	21613	3.3708416	0.9300618	0	33.0000000
bathrooms	21613	2.1147573	0.7701632	0	8.0000000
sqft_living	21613	2079.90	918.4408970	290.0000000	13540.00
sqft_lot	21613	15106.97	41420.51	520.0000000	1651359.00
floors	21613	1.4943090	0.5399889	1.0000000	3.5000000
waterfront	21613	0.0075418	0.0865172	0	1.0000000
view	21613	0.2343034	0.7663176	0	4.0000000
condition	21613	3.4094295	0.6507430	1.0000000	5.0000000
grade	21613	7.6568732	1.1754588	1.0000000	13.0000000
sqft_above	21613	1788.39	828.0909777	290.0000000	9410.00
year_built	21613	4.0936936	1.4829725	1.0000000	6.0000000
sqft_living15	21613	1986.55	685.3913043	399.0000000	6210.00
price	21613	540088.14	367127.20	75000.00	7700000.00

Identification of missing values

The MEANS Procedure

Variable	N Miss
price	0
sqft_lot	0
sqft_living15	0
bathrooms	0
bedrooms	0
condition	0
waterfront	0
view	0

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
4	bathrooms	Num	8	BEST12.	BEST32.
3	bedrooms	Num	8	BEST12.	BEST32.
10	condition	Num	8	BEST12.	BEST32.
1	date	Num	8	BEST12.	BEST32.
7	floors	Num	8	BEST12.	BEST32.
11	grade	Num	8	BEST12.	BEST32.
17	lat	Num	8	BEST12.	BEST32.
18	long	Num	8	BEST12.	BEST32.
2	price	Num	8	BEST12.	BEST32.
12	sqft_above	Num	8	BEST12.	BEST32.
13	sqft_basement	Num	8	BEST12.	BEST32.
5	sqft_living	Num	8	BEST12.	BEST32.
19	sqft_living15	Num	8	BEST12.	BEST32.
6	sqft_lot	Num	8	BEST12.	BEST32.
20	sqft_lot15	Num	8	BEST12.	BEST32.
9	view	Num	8	BEST12.	BEST32.
8	waterfront	Num	8	BEST12.	BEST32.
21	year_built	Num	8	BEST12.	BEST32.
22	year_renov	Num	8	BEST12.	BEST32.
14	yr_built	Num	8	BEST12.	BEST32.
15	yr_renovated	Num	8	BEST12.	BEST32.
16	zipcode	Num	8	BEST12.	BEST32.

3 Data Preprocessing

Data Pre-processing

- Removed the variables with less predictive power - Id, date, yr_renov, lat, long
- Converted zip code to continuous variable using the ZIPCITYDISTANCE function
- Normalized the dataset using z-score standardization (mean=0, std=1)
- Removed influential observations using cook's distance value
- Total independent variables : 15
- Total observations without influential: 20437

4 Multiple Linear Regression

Multiple Linear Regression

Used **stepwise** selection method

Equation: price_z = 0.199 - 0.050 x bedrooms_z + 0.059 x bathrooms_z + 0.221 x sqft_living_z + 0.055 x sqft_loft_z - 0.007 x floors_z + 0.153 x waterfront_z + 0.090 x view_z + 0.050 x condition_z + 0.225 x grade_z + 0.113 x sqft_above_z - 0.059 x year_built_z + 0.091 x sqft_living15_z - 0.278 x distance_z

R-square: 0.7838

Durbin-Watson D	1.996
Pr < DW	0.3961
Pr > DW	0.6039
Number of Observations	20437
1st Order Autocorrelation	0.002

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.19913	0.00955	20.85	<.0001	0
bedrooms_z	1	-0.05037	0.00306	-16.47	<.0001	1.71312
bathrooms_z	1	0.05961	0.00419	14.23	<.0001	3.11818
sqft_living_z	1	0.22076	0.00707	31.23	<.0001	7.88473
sqft_lot_z	1	0.05500	0.00328	16.79	<.0001	1.12233
floors_z	1	-0.00720	0.00324	-2.22	0.0264	2.14587
waterfront_z	1	0.15258	0.00607	25.13	<.0001	1.03851
view_z	1	0.09032	0.00291	31.00	<.0001	1.17362
condition_z	1	0.05024	0.00244	20.59	<.0001	1.19793
grade_z	1	0.22519	0.00425	52.96	<.0001	3.15202
sqft_above_z	1	0.11296	0.00624	18.10	<.0001	6.62900
year_built_z	1	-0.05973	0.00226	-26.48	<.0001	2.26839
sqft_living15_z	1	0.09066	0.00402	22.57	<.0001	2.94144
distance_z	1	-0.27755	0.00274	-101.44	<.0001	1.48763

Correlation Matrix

Pearson Correlation Coefficients, N = 20437 Prob > |r| under H0: Rho=0															
	bedrooms_z	bathrooms_z	sqft_living_z	sqft_lot_z	floors_z	waterfront_z	view_z	condition_z	grade_z	sqft_above_z	sqft_basement_z	year_built_z	sqft_living15_z	sqft_lot15_z	distance_z
bedrooms_z	1.00000	0.50788 <.0001	0.59937 <.0001	0.04184 <.0001	0.16328 <.0001	0.00322 0.6449	0.06726 <.0001	0.02856 <.0001	0.34529 <.0001	0.47758 <.0001	0.29678 <.0001	0.15795 <.0001	0.39918 <.0001	0.03815 <.0001	0.09062 <.0001
bathrooms_z	0.50788 <.0001	1.00000	0.72905 <.0001	0.08190 <.0001	0.50935 <.0001	0.03316 <.0001	0.13026 <.0001	-0.13468 <.0001	0.63845 <.0001	0.65200 <.0001	0.23108 <.0001	0.53557 <.0001	0.55824 <.0001	0.07639 <.0001	0.11561 <.0001
sqft_living_z	0.59937 <.0001	0.72905 <.0001	1.00000	0.19044 <.0001	0.35047 <.0001	0.03877 <.0001	0.21137 <.0001	-0.06715 <.0001	0.73788 <.0001	0.86190 <.0001	0.37619 <.0001	0.33905 <.0001	0.76499 <.0001	0.18195 <.0001	0.11188 <.0001
sqft_lot_z	0.04184 <.0001	0.08190 <.0001	0.19044 <.0001	1.00000	-0.02324 0.0009	0.00095 0.8922	0.04796 <.0001	-0.01189 0.0893	0.12342 <.0001	0.19948 <.0001	0.00707 0.3121	0.06031 <.0001	0.17968 <.0001	0.75643 <.0001	0.23600 <.0001
floors_z	0.16328 <.0001	0.50935 <.0001	0.35047 <.0001	-0.02324 0.0009	1.00000	0.02085 0.0029	-0.00140 0.8418	-0.27621 <.0001	0.46159 <.0001	0.53084 <.0001	-0.28618 <.0001	0.51413 <.0001	0.28075 <.0001	-0.02395 0.0006	0.00649 0.3534
waterfront_z	0.00322 0.6449	0.03316 <.0001	0.03877 <.0001	0.00095 0.8922	0.02085 0.0029	1.00000	0.18819 <.0001	0.00821 0.2403	0.04482 <.0001	0.03385 <.0001	0.01378 0.0488	-0.00524 0.4535	0.04156 <.0001	-0.00038 0.9561	-0.00686 0.3265
view_z	0.06726 <.0001	0.13026 <.0001	0.21137 <.0001	0.04796 <.0001	-0.00140 0.8418	0.18819 <.0001	1.00000	0.04282 <.0001	0.19036 <.0001	0.09904 <.0001	0.23142 <.0001	-0.05845 <.0001	0.22692 <.0001	0.04249 <.0001	-0.05628 <.0001
condition_z	0.02856 <.0001	-0.13468 <.0001	-0.06715 <.0001	-0.01189 0.0893	-0.27621 <.0001	0.00821 0.2403	0.04282 <.0001	1.00000	-0.16283 <.0001	-0.17497 <.0001	0.18870 <.0001	-0.36331 <.0001	-0.10984 <.0001	0.00208 0.7664	-0.07638 <.0001
grade_z	0.34529 <.0001	0.63845 <.0001	0.73788 <.0001	0.12342 <.0001	0.46159 <.0001	0.04482 <.0001	0.19036 <.0001	-0.16283 <.0001	1.00000	0.73604 <.0001	0.09474 <.0001	0.46039 <.0001	0.70279 <.0001	0.11787 <.0001	0.01929 0.0058
sqft_above_z	0.47758 <.0001	0.65200 <.0001	0.86190 <.0001	0.19948 <.0001	0.53084 <.0001	0.03385 <.0001	0.09904 <.0001	-0.17497 <.0001	0.73604 <.0001	1.00000	-0.14559 <.0001	0.44974 <.0001	0.74119 <.0001	0.19277 <.0001	0.22552 <.0001
sqft_basement_z	0.29678 <.0001	0.23108 <.0001	0.37619 <.0001	0.00707 0.3121	-0.28618 <.0001	0.01378 0.0488	0.23142 <.0001	0.18870 <.0001	0.09474 <.0001	-0.14559 <.0001	1.00000	-0.16026 <.0001	0.13822 <.0001	0.00278 0.6912	-0.19379 <.0001
year_built_z	0.15795 <.0001	0.53557 <.0001	0.33905 <.0001	0.06031 <.0001	0.51413 <.0001	-0.00524 0.4535	-0.05845 <.0001	-0.36331 <.0001	0.46039 <.0001	0.44974 <.0001	-0.16026 <.0001	1.00000	0.34319 <.0001	0.06534 <.0001	0.36907 <.0001
sqft_living15_z	0.39918 <.0001	0.55824 <.0001	0.76499 <.0001	0.17968 <.0001	0.28075 <.0001	0.04156 <.0001	0.22692 <.0001	-0.10984 <.0001	0.70279 <.0001	0.74119 <.0001	0.13822 <.0001	0.34319 <.0001	1.00000	0.19692 <.0001	0.15229 <.0001
sqft_lot15_z	0.03815 <.0001	0.07639 <.0001	0.18195 <.0001	0.75643 <.0001	-0.02395 0.0006	-0.00038 0.9561	0.04249 <.0001	0.00208 0.7664	0.11787 <.0001	0.19277 <.0001	0.00278 0.6912	0.06534 <.0001	0.19692 <.0001	1.00000	0.24615 <.0001
distance_z	0.09062 <.0001	0.11561 <.0001	0.11188 <.0001	0.23600 <.0001	0.00649 0.3534	-0.00686 0.3265	-0.05628 <.0001	-0.07638 <.0001	0.01929 0.0058	0.22552 <.0001	-0.19379 <.0001	0.36907 <.0001	0.15229 <.0001	0.24615 <.0001	1.00000

Principal Component Analysis

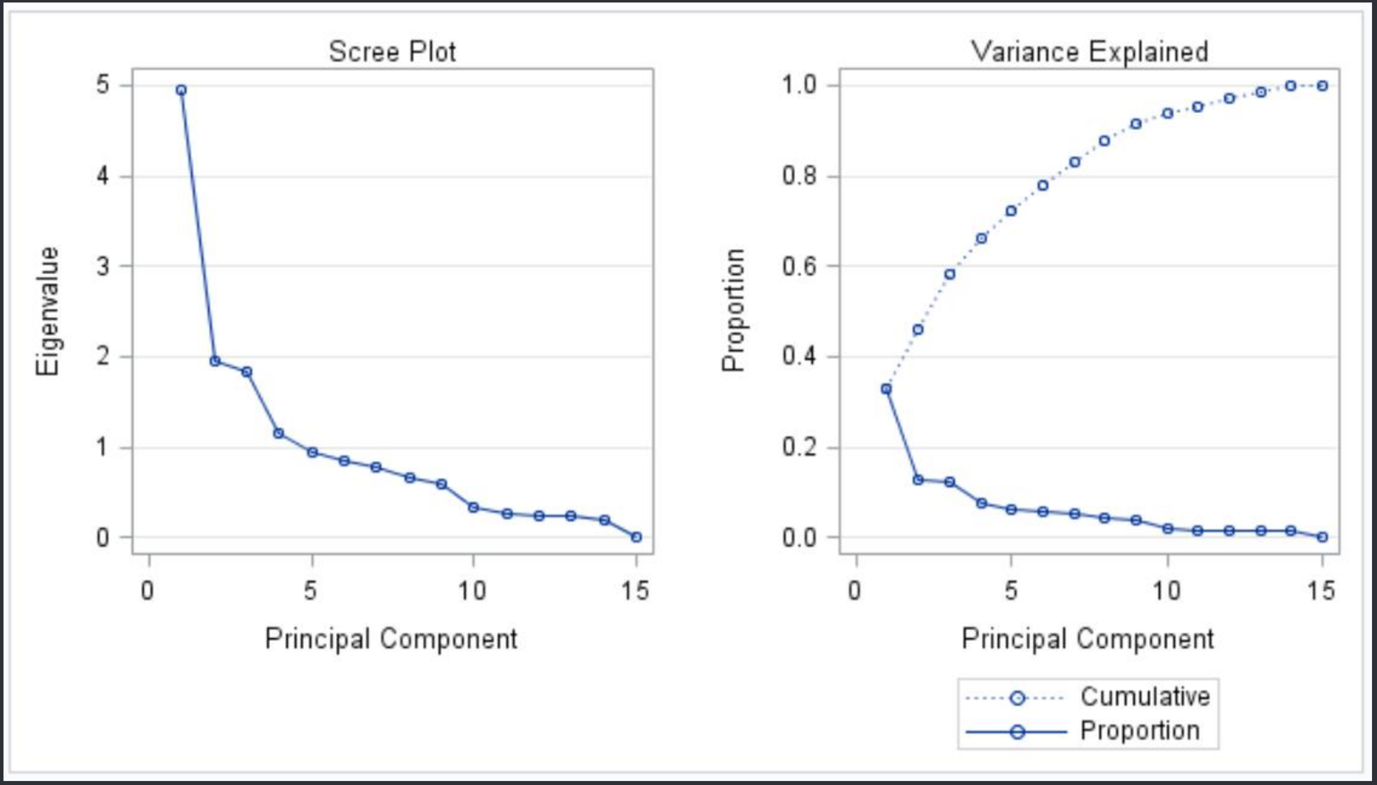
Implemented PCA to tackle the problem of correlation between independent variables and dimension reduction

Eigenvectors															
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15
bedrooms_z	0.257643	0.246972	-.073945	-.304662	0.256819	0.046910	0.252401	-.163240	0.644707	-.387395	0.069336	0.161953	0.011167	0.137645	0.000000
bathrooms_z	0.372013	0.043983	-.125009	-.080081	0.053374	-.088511	0.152182	0.351205	0.014386	0.138326	-.210717	-.704023	-.144456	0.316523	0.000000
sqft_living_z	0.405838	0.218363	-.016489	-.080114	-.006085	0.022075	0.015572	-.126580	-.066773	0.268190	-.080865	0.048031	-.010352	-.459844	-.683622
sqft_lot_z	0.109372	0.009076	0.635959	0.047379	-.209469	-.014430	0.158985	0.092236	0.077930	-.020453	-.463471	0.085725	0.513023	0.107515	0.000000
floors_z	0.255115	-.330959	-.211404	0.119044	-.229340	0.190352	0.105001	0.408721	0.290092	0.408845	0.282487	0.357333	0.146881	0.138718	0.000000
waterfront_z	0.023749	0.083319	-.015671	0.701645	0.417926	0.202782	0.512225	-.087485	-.117906	0.002312	0.010175	0.008447	0.006813	0.001886	0.000000
view_z	0.087291	0.306290	0.000732	0.555476	-.019340	-.248888	-.580253	0.183063	0.378074	-.093166	-.028206	-.030317	-.012540	-.063209	0.000000
condition_z	-.102191	0.368364	0.078794	-.163358	0.190243	0.681967	-.211042	0.477259	-.161785	-.117551	0.029024	0.026209	0.060657	-.046729	0.000000
grade_z	0.375549	0.026164	-.087255	0.088358	-.197097	0.068679	-.088296	-.039344	-.325058	-.252075	-.333085	0.437625	-.434705	0.359604	0.000000
sqft_above_z	0.404586	-.077562	0.003545	-.006730	-.031007	0.275046	-.103409	-.252037	0.059708	0.142411	-.152591	-.098245	-.044844	-.459047	0.640228
sqft_basement_z	0.052550	0.567763	-.038649	-.144011	0.044785	-.459500	0.219332	0.213560	-.239379	0.263042	0.121043	0.273227	0.061743	-.058406	0.350384
year_built_z	0.268424	-.371739	-.054957	-.019026	0.205105	-.273135	0.026406	0.390401	-.225842	-.539127	0.101851	0.040769	0.189607	-.356294	0.000000
sqft_living15_z	0.364416	0.111679	0.029839	0.041045	-.038374	0.066725	-.225856	-.341337	-.294201	-.071513	0.500381	-.146920	0.443441	0.344141	0.000000
sqft_lot15_z	0.108333	0.006745	0.639983	0.042211	-.195530	0.000003	0.141996	0.097260	0.048272	-.070956	0.486009	-.089306	-.500606	-.082782	0.000000
distance_z	0.103804	-.252114	0.316913	-.134046	0.707014	-.119701	-.307436	-.020533	-.005024	0.339069	-.045445	0.179051	-.110737	0.191690	0.000000

Principal Component Analysis

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.96416786	3.01380255	0.3309	0.3309
2	1.95036532	0.10419722	0.1300	0.4610
3	1.84616810	0.69874740	0.1231	0.5840
4	1.14742070	0.19865246	0.0765	0.6605
5	0.94876824	0.08746510	0.0633	0.7238
6	0.86130313	0.08392895	0.0574	0.7812
7	0.77737419	0.11702228	0.0518	0.8330
8	0.66035191	0.07627085	0.0440	0.8771
9	0.58408106	0.25689533	0.0389	0.9160
10	0.32718573	0.06995638	0.0218	0.9378
11	0.25722936	0.01186368	0.0171	0.9550
12	0.24536568	0.01320341	0.0164	0.9713
13	0.23216227	0.03410580	0.0155	0.9868
14	0.19805646	0.19805646	0.0132	1.0000
15	0.00000000		0.0000	1.0000

Principal Component Analysis



We chose first 10 principal components having cumulative of 0.9378 .

Multiple Regression using PCA

Durbin-Watson D	1.996
Pr < DW	0.3803
Pr > DW	0.6197
Number of Observations	20437
1st Order Autocorrelation	0.002

Equation:

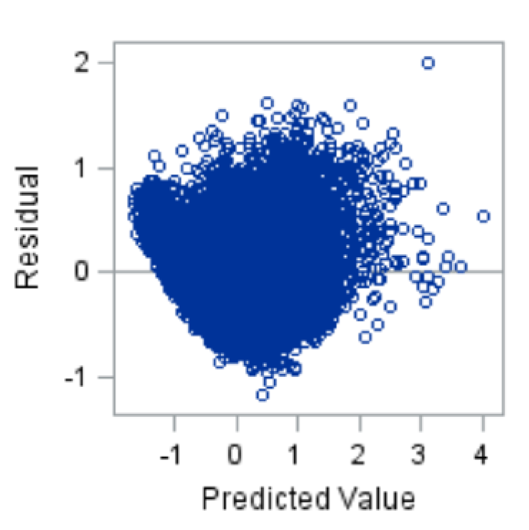
price_z = -0.118 + 0.191*Prin1 + 0.189*Prin2 - 0.073*Prin3 + 0.128*Prin4 - 0.244*Prin5 + 0.127*Prin6 + 0.014*Prin7 - 0.056*Prin8 - 0.094*Prin9 - 0.028*Prin10

R square : 0.7778

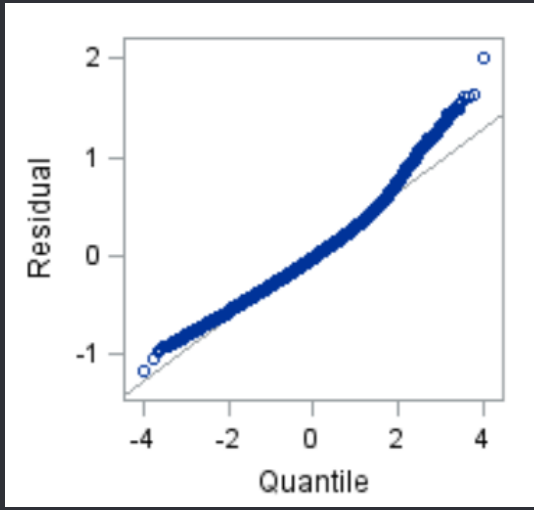
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.11790	0.00224	-52.72	<.0001	0
Prin1	1	0.19081	0.00100	190.08	<.0001	1.00000
Prin2	1	0.18991	0.00160	118.58	<.0001	1.00000
Prin3	1	-0.07337	0.00165	-44.57	<.0001	1.00000
Prin4	1	0.12769	0.00209	61.16	<.0001	1.00000
Prin5	1	-0.24407	0.00230	-106.30	<.0001	1.00000
Prin6	1	0.12707	0.00241	52.73	<.0001	1.00000
Prin7	1	0.01403	0.00254	5.53	<.0001	1.00000
Prin8	1	-0.05598	0.00275	-20.34	<.0001	1.00000
Prin9	1	-0.09355	0.00293	-31.97	<.0001	1.00000
Prin10	1	-0.02758	0.00391	-7.05	<.0001	1.00000

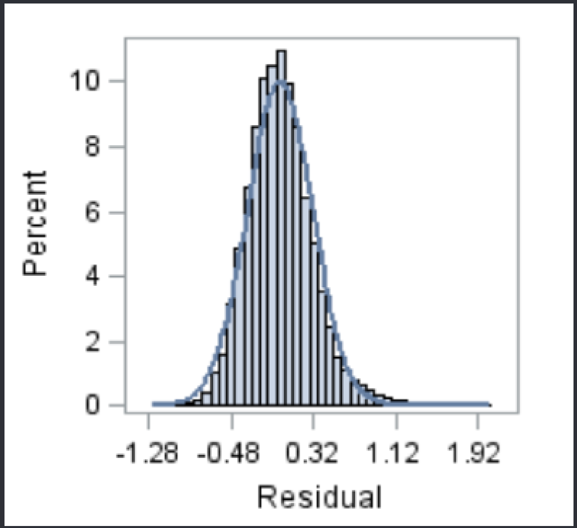
Residual fit plots



Scatterplot of the predicted value vs residuals.



QQPlot along the regression line with some distortion towards the end. But overall the plot is good.



Residuals are normally distributed without any skewness.

5

Logistic Regression

Logistic Regression

We applied logistic regression to classify the house prices into two categories:

1. Expensive - Houses having price greater and equal to the average price.

2. Not Expensive - Houses having price less than the average price.

Used first 10 principal components as independent variables to classify the house to be expensive or not.

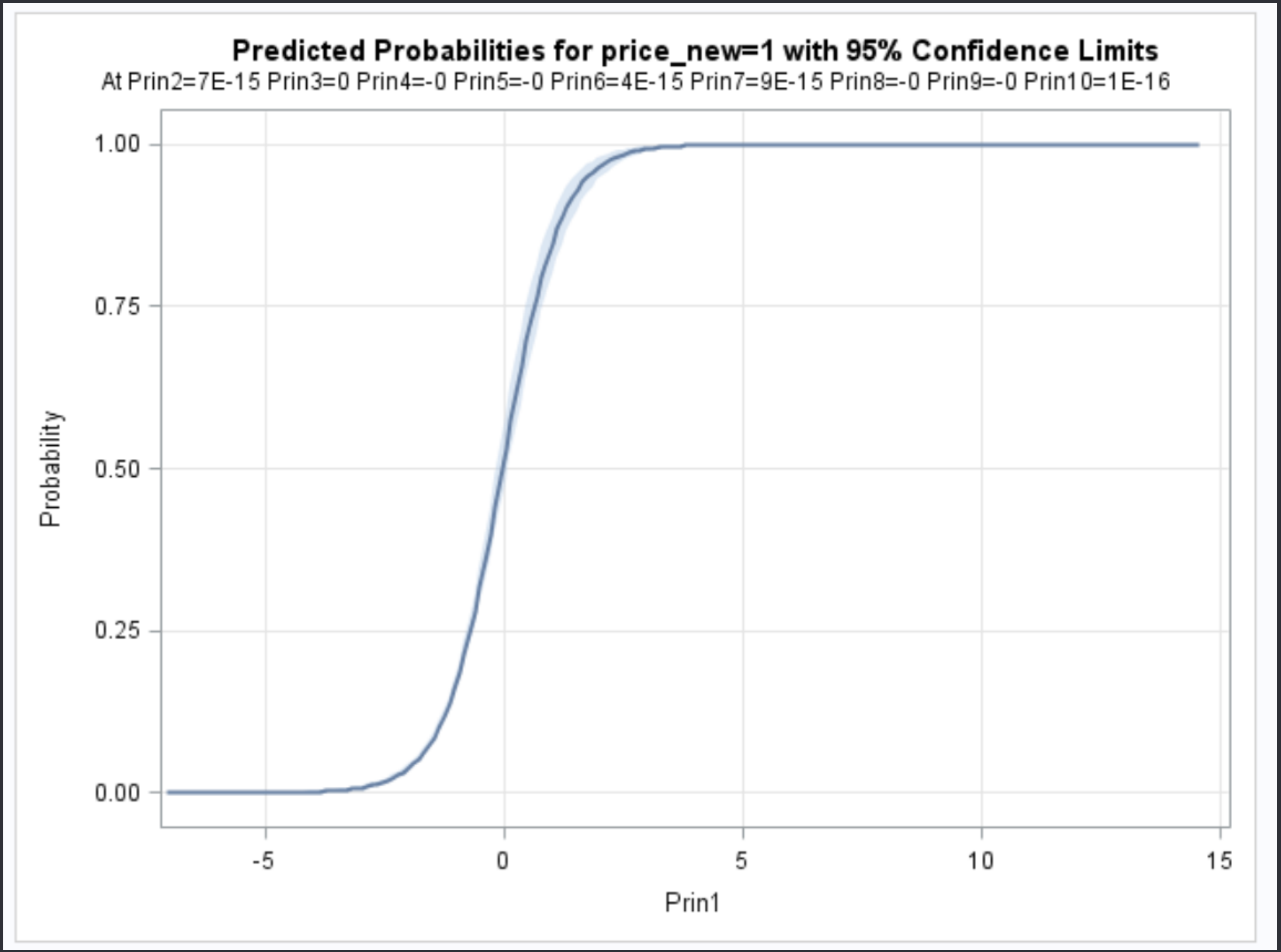
Logistic Regression

Logit(log of odds) =
0.058 + 1.651*Prin1 +
0.299*Prin2 -
0.080*Prin3 +
16.347*Prin4 +
7.442*Prin5 +
5.367*Prin6 _11.669*Prin7
- 2.418* Prin8 -
3.275*Prin9 - 0.511*Prin10

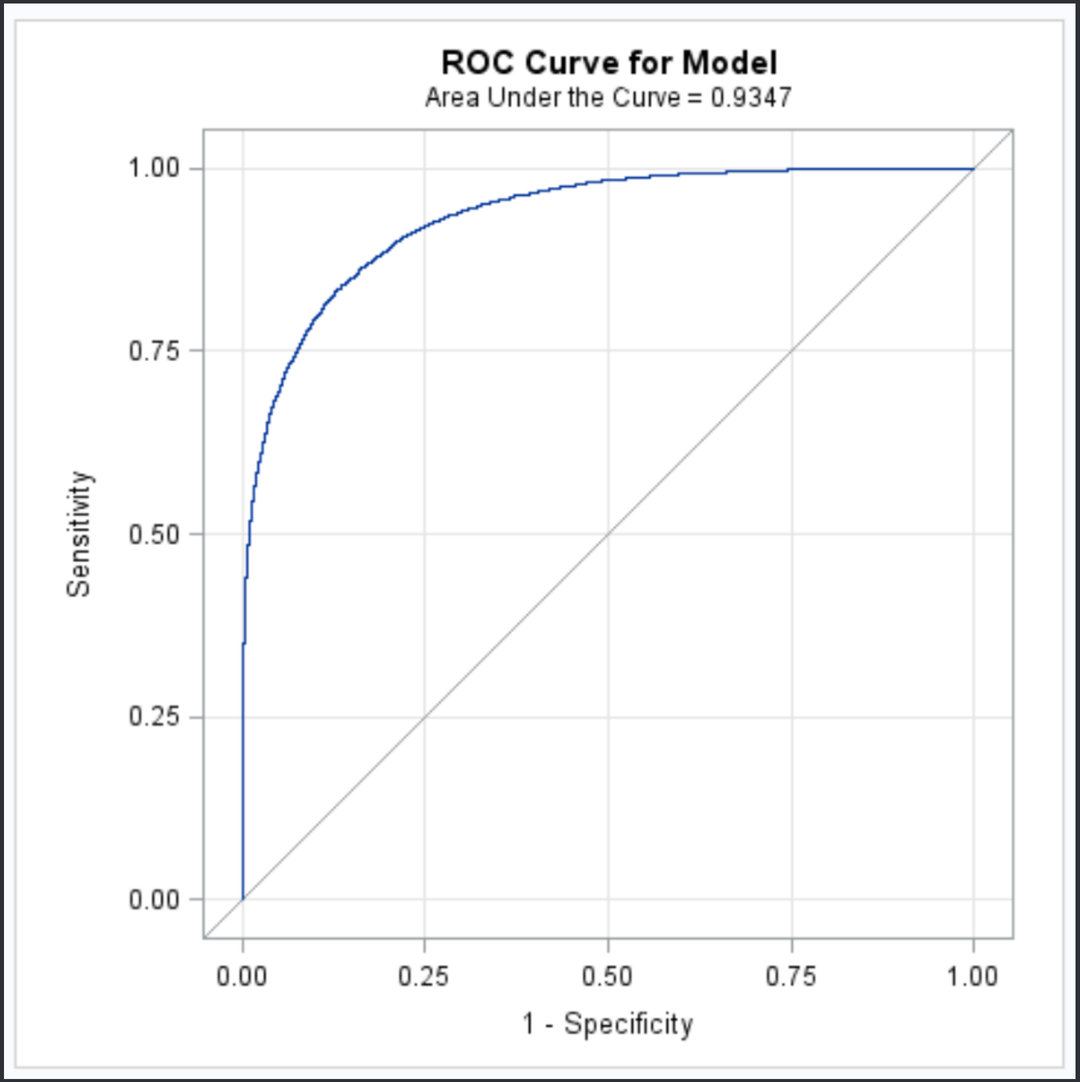
Odds = π(x) / [1-π(x)]

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0585	0.1007	0.3372	0.5615
Prin1	1	1.6516	0.0766	464.6588	<.0001
Prin2	1	2.9905	0.2557	136.7376	<.0001
Prin3	1	-0.8003	0.0516	240.4286	<.0001
Prin4	1	16.3470	2.1392	58.3927	<.0001
Prin5	1	7.4419	1.2702	34.3280	<.0001
Prin6	1	5.3667	0.6195	75.0388	<.0001
Prin7	1	11.6686	1.5615	55.8434	<.0001
Prin8	1	-2.4182	0.2681	81.3287	<.0001
Prin9	1	-3.2746	0.3607	82.4038	<.0001
Prin10	1	-0.5105	0.0425	144.0697	<.0001

Logistic Regression



Logistic Regression



c-stats: 0.935

6 Discriminant Analysis

Discriminant Analysis

- Statistical model to assess the adequacy of classification.
- DISCRIM procedure develops a discriminant criterion to classify each observation into one of the groups.
- PROC DISCRIM evaluates the performance of a discriminant criterion by estimating error rates (probabilities of misclassification) in the classification of future observations.
- Requires prior knowledge of the classes, usually in the form of a sample from each class.

Error Count Estimates for price_new

	0	1	Total
Rate	0.1229	0.1826	0.1528
Priors	0.5000	0.5000	

Number of Observations and Percent Classified into price_new

From price_new	0	1	Total
0	10596 87.71	1485 12.29	12081 100.00
1	1526 18.26	6830 81.74	8356 100.00
Total	12122 59.31	8315 40.69	20437 100.00
Priors	0.5	0.5	

Conclusion & Future Work

- Implement more classification techniques like Naive Bayes, Support Vector Machine and Artificial Neural Networks to make a model more robust and increase accuracy

- Extend our model to predict the price of houses in more county's across Washington state and USA

- Collect data about people's opinion on the house when they visit and implement sentiment analysis on their opinion. This will not be a measure in predicting the price but help real estate agents to make necessary improvements to the house, especially for houses that are not sold.

Thanks!

ANY QUESTIONS?

Special thanks to professor Khasha Dehand for his teachings and continual support during the entire course.