DOCUMENTATION OF DATA PROCESSING STEPS

- Identify and separate the categorical from the numerical data.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	price	crime_rate	resid_area	air_qual	room_num	age	dist1	dist2	dist3	dist4	teachers	poor_prop	n_hos_beds	n_hot_rooms	rainfall	parks	airport	waterbody	bus_ter
	24	0.00632	32.31	0.538	6.575	65.2	4.35	3.81	4.18	4.01	24.7	4.98	5.48	11.192	23	0.049347306	YES	River	YES
	21.6	0.02731	37.07	0.469	6.421	78.9	4.99	4.7	5.12	5.06	22.2	9.14	7.332	12.1728	42	0.046145633	NO	Lake	YES
	34.7	0.02729	37.07	0.469	7.185	61.1	5.03	4.86	5.01	4.97	22.2	4.03	7.394	101.12	38	0.045763966	NO	None	YES
	33.4	0.03237	32.18	0.458	6.998	45.8	6.21	5.93	6.16	5.96	21.3	2.94	9.268	11.2672	45	0.047150598	YES	Lake	YES
	36.2	0.06905	32.18	0.458	7.147	54.2	6.16	5.86	6.37	5.86	21.3	5.33	8.824	11.2896	55	0.039474005	NO	Lake	YES
	28.7	0.02985	32.18	0.458	6.43	58.7	6.22	5.8	6.23	5.99	21.3	5.21	7.174	14.2296	53	0.045909647	YES	None	YES
	22.9	0.08829	37.87	0.524	6.012	66.6	5.87	5.47	5.7	5.2	24.8	12.43	6.958	12.1832	41	0.052169908	YES	River	YES
	22.1	0.14455	37.87	0.524	6.172	96.1	6.04	5.85	6.25	5.66	24.8	19.15	5.842	12.1768	56	0.057074901	NO	Lake	YES
	16.5	0.21124	37.87	0.524	5.631	100	6.18	5.85	6.3	6	24.8	29.93	5.93	12.132	55	0.056302495	YES	None	YES
	18.9	0.17004	37.87	0.524	6.004	85.9	6.67	6.55	6.85	6.29	24.8	17.1	9.478	14.1512	45	0.050727252	YES	River	YES
	15	0.22489	37.87	0.524	6.377	94.3	6.65	6.31	6.55	5.88	24.8	20.45	6	11.12	29	0.057775258	NO	Lake	YES
	18.9	0.11747	37.87	0.524	6.009	82.9	6.27	5.93	6.51	6.19	24.8	13.27	9.278	13.1512	23	0.055236508	NO	Lake and River	YES
	21.7	0.09378	37.87	0.524	5.889	39	5.76	5.14	5.58	5.33	24.8	15.71	5.534	10.1736	57	0.057423248	YES	Lake and River	YES
	20.4	0.62976	38.14	0.538	5.949	61.8	4.72	4.59	4.93	4.59	19	8.26	5.908	14.1632	39	0.053463955	YES	None	YES
	18.2	0.63796	38.14	0.538	6.096	84.5	4.6	4.2	4.48	4.58	19	10.26	6.964	13.1456	49	0.059882129	NO	None	YES
	19.9	0.62739	38.14	0.538	5.834	56.5	4.6	4.35	4.72	4.32	19	8.47	8.498	14.1592	28	0.059750758	YES	River	YES
	23.1	1.05393	38.14	0.538	5.935	29.3	4.66	4.39	4.52	4.43	19	6.58	5.462	10.1848	46	0.054698587	NO	None	YES
	17.5	0.7842	38.14	0.538	5.99	81.7	4.56	4.15	4.36	3.97	19	14.67	5.45	11.14	56	0.05478547	NO	Lake	YES
	20.2	0.80271	38.14	0.538	5.456	36.6	3.8	3.52	3.86	4	19	11.69	8.504	12.1616	41	0.054250839	YES	Lake and River	YES
	18.2	0.7258	38.14	0.538	5.727	69.5	3.98	3.65	4	3.57	19	11.28	8.564	12.1456	27	0.057770199	NO	Lake and River	YES
	13.6	1.25179	38.14	0.538	5.57	98.1	3.93	3.59	4.09	3.58	19	21.02	8.272	15.1088	44	0.048317527	YES	Lake and River	YES
	19.6	0.85204	38.14	0.538	5.965	89.2	4.11	3.72	4.34	3.88	19	13.83	9.192	14.1568	23	0.054040555	YES	None	YES
	15.2	1.23247	38.14	0.538	6.142	91.7	4.18	3.98	4.31	3.45	19	18.72	5.804	14.1216	48	0.057413663	YES	River	YES

Feature Engineering

Refer to DATA PREPROCESSING sheet for follow up

Missing values

From our EDD, we notice the counts of the ***n_hos_beds*** is not equal to 506 which is the number of observations. This shows that there are missing values in that feature, and to take care of this, we take the average of value (7.899767068) of the column feature and fill in the missing gaps with that value.

Outliers

From our descriptive statistics shown on EDD sheet, there is a huge deviation shown in the mean and median, minimum and 25th percentile (smallest 25) value, maximum and 75th percentile (largest 125) value in the ***crime_rate*** column feature. All these signs are indications of outliers. So we proceed to replacing all values of the feature greater than the 95th percentile (15.78915) with a value equals to 2 * 95th percentile.

Variable transformation

There is sort of redundancy in the dist1, dist2, dist3 and dist4 column features. To handle this, we create a new feature, ***av_dist***, by taking the average of these features and removing them afterwards.

Feature encoding

Since the regression model cannot handle text or strings format, we have to encode the categorical features. We do this by using the IF() function, where set our value to 1 if true and 0 if false.

airport	Lake	River	Lake and River	bus_ter
=IF($N3="YES", 1, 0)	=IF($Q3="Lake", 1, 0)	=IF($Q3="River", 1, 0)	=IF($Q3="Lake and River", 1, 0)	=IF($S3="YES", 1, 0)
=IF($N4="YES", 1, 0)	=IF($Q4="Lake", 1, 0)	=IF($Q4="River", 1, 0)	=IF($Q4="Lake and River", 1, 0)	=IF($S4="YES", 1, 0)
=IF($N5="YES", 1, 0)	=IF($Q5="Lake", 1, 0)	=IF($Q5="River", 1, 0)	=IF($Q5="Lake and River", 1, 0)	=IF($S5="YES", 1, 0)
=IF($N6="YES", 1, 0)	=IF($Q6="Lake", 1, 0)	=IF($Q6="River", 1, 0)	=IF($Q6="Lake and River", 1, 0)	=IF($S6="YES", 1, 0)

Feature Selection

After the correlation analysis was done on the CORR sheet, there seems to be a high positive correlation between ***air_qual*** and ***parks***. This high correlation can disrupt our model from learning properly, so we have to remove one out of the two features. Our preference goes to the feature with a higher correlation to the target variable and this is ***air_qual*** so we remove the ***parks*** feature.

The data is now ready for use. (see PREPROCESSED DATA sheet)

INFERENCES

Upon correlation analysis, the two most correlated features to the target variable (price) are room_num and poor_prop. With room_num having a positive correlation and prop_prop, a strong negative correlation. Three regression models were therefore built with price as target variable, they are;

1. Simple linear regression on room_num feature.
2. Simple linear regression on poor_prop feature.
3. Multi regression on all features.

(1) Simple linear regression on room_num feature; The r-square value of this model found under the summary statistics of the Linear Reg 1 sheet is seen to be 0.484838974, signifying that about 48.48% of the changes in the data is explained by our model. The standard error is 6.597015858, signifying that our predicted values of price by the model has an error of +/- 6.597015858 on average. Also, on our Anova table, the F value is far greater than the F critical value meaning our regression model is significant.

(2) Simple linear regression on poor_prop; The r-square value of this model found under the summary statistics of the Linear Reg 2 sheet is seen to be 0.548837968, signifying that about 54.88% of the changes in the data is explained by our model. The standard error is 6.1736542, signifying that our predicted values of price by the model has an error of +/- 6.1736542 on

average. The F value is greater than the F critical value, signifying that the regression model is significant.

(3) Multi regression on all features; The r-square value of this model found under the summary statistics of the Multi Reg sheet is seen to be 0.720994242, signifying that about 72.10% of the changes in the data is explained by our model. The standard error is 4.923793596, signifying that our predicted values of price by the model has an error of +/- 4.923793596 on average. Also, on our Anova table, the F value is far greater than the F critical value meaning our regression model is significant.

The Multi Regression model seems to be the most effective since its R square value is larger and the standard error is the most minimal. This is accompanied by the Linear regression model on the poor_prop feature and lastly the room_num feature.