



INFOSYS 722 Research Report

Factors Affecting Sydney House Prices

Iteration 2 ISAS (Steps 1 – 8)

Shiyu Lin

UPI: slin648

ID: 502522556

Department of Information Systems and Operations Management
(ISOM)

University of Auckland, New Zealand

Contents

1. Situation understanding	1
1.1 Identify the objectives of the situation	1
1.2 Assess the situation	1
1.3 Determine data mining objectives	2
1.4 Produce a project plan.....	2
2. Data understanding	4
2.1 Collect initial data	4
2.2 Describe the data.....	5
2.3 Explore the data.....	7
2.3.1 Import the Data	7
2.3.2 Explore and visualize	8
2.4 Verify the data quality	10
3. Data preparation	12
3.1 Select the data	12
3.2 Clean the data.....	13
3.3 Construct the data	15
3.4 Integrate various data sources.....	16
3.5 Format the data as required	17
4. Data transformation	19
4.1 Reduce the data.....	19
4.2 Project the data	20
5. Data-mining algorithm(s) selection	22
5.1 Match and discuss the objectives of data mining to data mining methods.....	22
5.2 Select the appropriate data-mining method(s) based on discussion.....	22
6. Data-mining algorithm(s) selection	24
6.1 Conduct exploratory analysis and discuss	24

6.2 Select data-mining algorithms based on discussion.....	24
6.3 Build/Select appropriate model(s) and choose relevant parameter(s).....	25
6.3.1 Linear regression model	25
6.3.2 Multivariate Regression model.....	26
6.3.3 Random Forest model	27
7. Data Mining	29
7.1 Create and justify test designs.....	29
7.2 Conduct data mining – classify, regress, cluster, etc.	29
7.3 Search for patterns	31
8. Interpretation	33
8.1 Study and discuss the mined patterns.....	33
8.2 Visualize the data, results, models, and patterns.....	33
8.3 Interpret the results, models, and patterns	37
8.4 Assess and evaluate results, models, and patterns	39
8.5 Iterate prior steps (1 – 7) as required	39
References	41

1. Situation understanding

1.1 Identify the objectives of the situation

Sustainable cities and human settlements is one of the 17 Sustainable Development Goals of the UN. [1] Cities are centers of business, culture, science, productivity, social, human and economic development. Urban planning, transport systems, water, sanitation, waste management, disaster risk reduction, access to information, education and capacity building are all issues relevant to sustainable urban development.

In 2008, the global urban population exceeded the rural population for the first time in history. [1] This milestone marks the arrival of a new “urban millennium”, where by 2050 two-thirds of the world’s population is expected to live in urban areas.

In some cities, housing prices have experienced severe housing bubbles, such as Tokyo, where house prices were as high as \$220,000 per square meter, and now are far below this value. Today, housing prices in some cities are seriously too high for local income levels, such as Shenzhen and Shanghai in China. All this poses a challenge to achieve the goal of sustainable cities and human settlements.

"Promoting sustainable human settlements development" is the subject of Chapter 7 of Agenda 21, which calls for providing adequate shelter for all, and providing adequate environment and support.

We know that it is very difficult to provide shelter for everyone while ensuring that the environment of the shelter is good. The price of a house is affected by many factors. This includes internal factors, such as the size of the house, the size of the house, etc.; it also includes external factors, such as the location of the house, the crime rate in the neighborhood where the house is located, etc. [2] A report by Etch Real Estate states that one of the most important short-term factors affecting house prices is interest rates. When the cost of borrowing money decreases, the number of qualified buyers increases. The price of a house is also related to the macroeconomic situation in the area.

1.2 Assess the situation

In the current situation, we need to collect data and analyze the data to report the factors affecting house prices, and our recommendations. This conclusion and recommendations are not based on experience - in fact I am not an expert in business or real estate. We need to analyze the data in the manner of a data scientist and report our findings.

The data we are looking for should contain one or more response variables to record house

prices; it should also contain multiple explanatory variables that explain what factors affect house prices. Each explanatory variable can be a potentially significant factor, but it can also be a trivial factor. Importantly, the data needs to be relevant: we need a row of data to describe a particular property transaction, or the price of a particular property.

Since we are not allowed to gather our own data without having Ethics Approval, and cannot carry out interviews or perform surveys without applying for Ethics Approval from the University of Auckland, I need to look for publicly available datasets. These datasets are provided and made available to the public by some agencies, who have ensured that the datasets do not contain sensitive information, such as other people's privacy and personal information.

The data we collect may contain some quality issues, such as outliers, or incomplete data. We need to examine and process the data to make sure it is reliable. This allows us to build reliable models and draw meaningful conclusions.

There are many factors that affect house prices, some of which may not be described in the explanatory variables of the dataset. We assume that factors not described in the dataset are trivial, but in fact it may not be the case.

To complete this iteration, we are required to use IBM SPSS Modeler to do all the work. I'll be doing some data collection/splitting in R and the all rest in IBM SPSS Modeler.

1.3 Determine data mining objectives

Our objectives are Explanation and Prediction.

Explanation: I'll make assertions such as "a 1% increase in neighborhood crime, a 2% decrease in home prices, *ceteris paribus*". I will also build models to find out which factors have the greatest impact on house prices.

Prediction: I will split the dataset into training and test sets, model on the training set, and test the accuracy of the model on the test set.

Desired Outcome:

1. Get to know what factors affect house prices the most, and what's trivial.
2. Make the model obtain a high prediction accuracy in the test set.
3. Based on the actual situation, I will explain my conclusions and put forward some logical conjectures.

1.4 Produce a project plan

Step	Duration	Resources	Risks
Situation understanding	3 days	references	insufficient progress
Technical preparations	3 days	Technical Support	Software cannot be installed
Data preprocessing	5 days	EXCEL R SPSS Modeler etc.	Too much data is lost
Modeling	6 days	R SPSS Modeler etc.	Not familiar with SPSS Modeler software
Evaluation	3 days	R SPSS Modeler etc.	Not familiar with SPSS Modeler software

Table 1.4.1: Project Plan

2. Data understanding

2.1 Collect initial data

For this research, the data came from the R package [3]HRW: Datasets, Functions and Scripts for Semiparametric Regression Supporting Harezlak, Ruppert & Wand (2018). This is an open-source R package that contains a large number of datasets.

First, I load the HRW package in R and load the SydneyRealEstate dataset. Then I export the dataset file in CSV format.

A1																		
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		logSalePrice	lotSize	longitude	latitude	saleDate	saleQtyr	inRate	postCode	crimeDen	crimeRate	income	distToBus	distToCoast	distToPark	distToRail	disToCoast	disToPark
2	1	14.00257076	645	151.0814	-33.8817	27/03/2002	0	0.131753	2135	658.072	3.929212	1564	0.153313	2.14268	6.232008	0.280569	1.14593	1.639267
3	2	13.02422334	662.925	150.9527	-33.8653	26/11/2000	4	0.08051	2165	369.1073	0.270408	571	0.075954	5.004402	6.828095	0.491962	0.874182	0.874715
4	3	13.73700649	812.0083	151.2883	-33.7693	21/11/2001	4	0.054411	2096	110.1008	1.29077	1109	0.013197	0.62769	5.015821	0.158354	10.10382	10.42105
5	4	12.06303006	553	150.815	-33.7326	01/06/2000	2	0.003506	2770	220.0644	0.20692	804	0.047612	19.48795	7.685946	0.274657	3.19051	3.18957
6	5	14.34314229	719	151.2041	-33.7817	25/09/2000	3	0.16782	2065	468.3177	1.18812	1444	0.236832	1.409012	3.996907	0.378831	0.986787	1.164581
7	6	12.38986839	750.5361	150.8981	-33.7202	20/02/2000	1	0.002329	2763	220.0644	0.20692	1285	0.179069	14.0219	4.623051	1.917114	1.351861	1.36085
8	7	12.06303006	575	150.8036	-33.7544	11/12/2000	4	0.023668	2763	220.0644	0.20692	504	0.122905	19.52995	9.820023	0.067935	1.353585	2.62654
9	8	12.99402982	700.2626	151.0421	-33.8653	13/07/2000	3	0.035158	2711	589.1128	0.251134	823	0.056877	1.443709	4.243691	0.190798	0.482116	0.728719
10	9	12.6549235	721	151.0355	-33.8045	19/10/2000	4	0.023668	2141	573.3682	0.333071	804	0.182562	3.406468	6.87696	0.713	0.726825	0.818422
11	10	12.98670099	584.632	151.0993	-33.9731	05/02/2000	1	0.042375	2220	473.294	0.147364	969	0.159443	1.348781	5.143559	0.087451	0.705762	0.705162
12	11	13.17391244	579.0058	151.0813	-33.9327	15/02/2000	1	0.017001	2196	465.689	0.140951	1016	0.027504	3.529289	4.791275	0.270779	1.496763	1.495848
13	12	15.05876233	927.0551	151.168	-33.8659	28/05/2000	2	0.071316	2039	1083.968	0.228337	1230	0.067356	0.670259	1.397773	0.669157	0.603387	0.786998
14	13	13.78711853	795.5046	151.0996	-34.0152	07/04/2000	2	-0.11335	2224	80.1769	1.32836	1104	0.072364	0.791808	4.792567	0.152579	2.277419	2.349885
15	14	13.35216607	654.9419	151.2726	-33.7585	10/08/2000	3	0.064422	2100	110.1008	1.29077	888	0.299278	2.204007	3.950385	0.174994	9.114079	9.316566
16	15	13.1467167	557.0994	151.1979	-33.9307	05/10/2000	4	0.038316	2029	390.811	0.24248	994	0.072094	1.847015	1.711343	0.271505	0.589373	1.289842
17	16	13.79107371	816	151.2114	-33.7873	21/09/2000	3	0.041093	2060	486.3177	1.18812	1406	0.011656	0.384682	1.673633	0.377183	0.288923	0.306036
18	17	14.50376549	758.7136	151.0953	-33.9987	27/09/2000	3	-0.11174	2224	80.2769	1.32836	1849	0.094697	10.059	5.346878	0.721449	2.078731	2.498404
19	18	14.4217852	1020.598	151.2047	-33.7762	26/07/2000	3	0.041093	2068	126.3732	1.07858	1643	0.599931	0.153423	0.344274	0.259352	0.258791	0.732825
20	19	13.51212532	957	151.2814	-33.7018	10/08/2000	3	0.113271	2101	84.943	1.14624	1196	0.069368	0.776442	3.671885	0.171891	13.1273	13.12643
21	20	13.36231304	777.0477	151.1073	-34.0361	20/06/2000	2	-0.00275	2234	80.2769	1.32836	1575	0.045589	3.21746	4.64155	0.636228	3.290786	3.587575
22	21	12.51907444	742.9089	150.8376	-33.7681	24/08/2000	3	0.039628	2766	220.0644	0.20692	841	0.309028	16.07044	4.18288	0.383044	0.502369	0.738415
23	22	12.98670099	971.8822	150.9991	-34.0522	09/07/2000	3	-0.01984	2233	80.2769	1.32836	1409	0.114695	4.73522	3.728193	0.063423	2.255975	2.251766
24	23	12.6338942	692.6026	151.2784	-33.7693	17/08/2000	3	0.166422	2100	110.1008	1.29077	980	0.073323	1.447253	4.197033	0.51718	9.278847	9.559357
25	24	12.55115553	600	150.8541	-34.0461	13/11/2000	4	0.025555	2560	116.2035	2.49854	971	0.099985	13.873	10.63154	0.183404	1.720151	2.252655
26	25	12.50008953	505	150.7413	-33.9454	11/04/2000	2	0.140969	2567	33.41461	1.535009	1330	0.187552	23.1701	13.65116	0.522073	2.821916	5.468294
27	26	12.6792162	806	150.7703	-33.0548	28/09/2000	3	0.050504	2759	90.44356	1.219255	1148	0.307205	21.65745	10.42508	0.352061	3.514822	5.322654
28	27	13.33154138	806.0218	151.0405	-33.9339	24/08/2000	3	0.015999	2126	37.79187	1.020721	1651	0.14697	9.045588	1.310153	0.083022	2.573548	2.941045
29	28	12.25105548	607	150.9055	-33.7371	10/01/2000	1	0.027326	2170	97.21553	1.933134	790	0.206831	5.568327	1.941259	0.390359	1.177883	1.549314
30	29	13.25582956	747.4111	151.1131	-33.9312	16/06/2000	2	0.021325	2206	545.6869	1.40951	946	0.040832	1.996628	2.745236	0.367329	0.584435	0.60925
31	30	13.19323671	440	151.1007	-33.8926	26/01/2000	1	0.007619	2136	981.4476	0.239182	1007	0.03198	1.189723	5.753582	0.370217	1.621108	1.645409
32	31	13.19800999	474	150.8953	-33.8724	28/11/2000	4	0.068339	2176	369.1073	0.270408	1096	0.137422	7.883884	3.270401	0.339917	4.665447	4.729717
33	32	12.50486281	827	151.051	-33.7137	16/10/2000	4	0.02174	2126	37.79187	1.020721	1710	0.068552	11.7075	20.27454	0.134307	3.221708	3.220978
34	33	13.33905051	822	150.8262	-33.7457	22/03/2000	1	-0.00613	2560	116.2035	2.49854	910	0.084523	17.9722	10.30639	0.351656	1.723222	1.722089
35	34	12.36176197	709	150.7122	-33.0723	20/12/2000	4	0.029987	2750	90.44356	1.219255	698	0.010429	2.279854	6.921143	0.109505	1.322054	1.684497

Figure2.1.1: Clean Dataset

Since this dataset is "clean", I added some missing values so that the data processing steps can follow.

For the `distToBusStop` variable, I randomly lost 20% of the data; for `distToCoastline`, I dropped more data with larger values. See section 2.4 for specific discarding rules.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		logSalePri	lotSize	longitude	latitude	saleDate	saleQtr	infRate	postCode	crimeDens	crimeRate	income	distToBust	distToCoa	distToNat	distToPark
2	1	14.00257	645	151.0814	-33.8817	23/07/200	3	0.131753	2135	658.072	0.329212	1564	0.153313	2.14268	6.923308	0.280569
3	2	13.02422	662.925	150.9527	-33.8653	26/11/200	4	0.08051	2165	369.1703	0.207048	571		5.004402	6.828095	0.491962
4	3	13.73701	812.0083	151.2883	-33.7693	21/11/200	4	0.054411	2096	110.1008	0.129077	1109	0.013197	0.62769	5.015821	0.158354
5	4	12.06303	553	150.815	-33.7326	01/06/200	2	0.003506	2770	220.0644	0.20692	804	0.047612		7.685946	0.274657
6	5	14.34314	719	151.2041	-33.817	25/09/200	3	0.016782	2065	486.3177	0.18812	1444	0.236832	1.409012	3.996907	0.378831
7	6	12.38886	750.5361	150.8981	-33.7202	20/02/200	1	0.002329	2763	220.0644	0.20692	1285	0.179069		4.623051	1.917114
8	7	12.06303	575	150.8036	-33.7544	11/12/200	4	0.082403	2770	220.0644	0.20692	504	0.122905		9.082003	0.067935
9	8	12.99403	700.2626	151.0421	-33.8045	13/07/200	3	0.035158	2117	589.1128	0.251134	823	0.056877	1.443709	4.243691	0.190798
10	9	12.65492	721	151.0355	-33.8653	19/10/200	4	0.023668	2141	573.3682	0.333071	804	0.128562	3.420364	6.876696	0.713
11	10	12.9867	584.632	151.0993	-33.9731	05/02/200	1	0.042375	2220	473.294	0.147364	969	0.159443	1.348787	5.143559	0.087451
12	11	13.17391	579.0058	151.0813	-33.9327	15/02/200	1	0.017001	2196	545.6869	0.140951	1016	0.027504	3.529289	4.792175	0.270779
13	12	15.05876	927.0551	151.168	-33.8659	28/05/200	2	0.107316	2039	1083.968	0.228337	1230	0.067356	0.670259	1.397773	0.669157
14	13	13.78712	795.5046	151.0996	-34.0152	27/04/200	2	-0.01135	2224	80.2769	0.132836	1104	0.072364	0.791808	4.792567	0.152579
15	14	13.35216	654.9419	151.2726	-33.7585	10/08/200	3	0.166422	2100	110.1008	0.129077	888	0.299278	2.204007	3.950385	0.174994
16	15	13.14672	557.0994	151.1979	-33.9307	05/10/200	4	0.038316	2020	390.8811	0.24248	994	0.072094	1.847015	7.171343	0.271505
17	16	13.79107	816	151.2114	-33.7873	21/09/200	3	0.041093	2069	486.3177	0.18812	1406	0.011656	0.384682	1.673633	0.377183
18	17	14.50377	758.7136	151.0953	-33.9987	27/09/200	3	-0.01174	2224	80.2769	0.132836	1849		0.067109	5.346878	0.721449
19	18	14.42179	1020.598	151.2047	-33.7762	26/07/200	3	0.041093	2069	126.3723	0.107858	1643	0.599931	0.153423	0.342474	0.253532
20	19	13.51213	957	151.2814	-33.7018	10/08/200	3	0.113271	2101	84.4943	0.14624	1196	0.069368	0.776442	3.671885	0.171891
21	20	13.36231	777.0477	151.0073	-34.0361	20/06/200	2	-0.00275	2234	80.2769	0.132836	1575	0.045589	3.281713	4.64155	0.636228
22	21	12.51905	742.9089	150.8376	-33.7681	24/08/200	3	0.039628	2766	220.0644	0.20692	841	0.309028		9.18828	0.383044
23	22	12.9867	971.8822	150.9991	-34.0522	09/07/200	3	-0.01984	2233	80.2769	0.132836	1409	0.114695	4.723052	3.728193	0.063423
24	23	12.63839	692.6026	151.2784	-33.7693	17/08/200	3	0.166422	2100	110.1008	0.129077	980	0.073323	1.447253	4.197033	0.51718
25	24	12.55116	600	150.8541	-34.0461	13/11/200	4	0.025555	2560	116.2035	0.249854	971	0.099985		10.63154	0.183404
26	25	12.50009	505	150.7413	-34.0544	11/04/200	2	0.140969	2567	33.41461	0.153509	1330	0.187552		13.65176	0.522073
27	26	12.67922	806	150.7703	-33.7935	28/09/200	3	0.050504	2759	90.44356	0.212955	1148	0.307205		10.42508	0.352061
28	27	13.33154	806.0218	151.0405	-33.7339	24/08/200	3	0.015999	2126	37.79187	0.120721	1651	0.14697	9.045588	1.310153	0.083202
29	28	12.25151	607	150.9055	-33.9371	10/01/200	1	0.027326	2170	97.21553	0.193314	709	0.206831	5.568327	1.941259	0.390359
30	29	13.25583	747.4111	151.1131	-33.932	26/06/200	2	0.031235	2206	545.6869	0.140951	946	0.040832	1.996628	7.245236	0.367329
31	30	13.19324	440	151.1007	-33.8916	16/01/200	1	0.007619	2136	981.4476	0.239182	1007		1.189723	5.753582	0.370217

Figure2.1.2: Dataset with missing values

2.2 Describe the data

The data has 37676 rows, representing 37676 transaction records. Each record represents a unique set of properties. Meanwhile, the data has 39 columns, each of which represents an attribute. That said, there are 39 different underlying factors that affect property prices.

The data is in csv format. I purposely export the data to csv format for later use.

[4]Specifically, this data frame contains the following columns:

--logSalePrice: The natural logarithm of the property price. In log(AUD).

--lotSize: The size of the property. In square meters.

--longitude: The longitude of the property's location.

--latitude: The latitude of the property's location.

--saleDate: The sale date of the property.

--saleQtr: The sale quarter of the property.

--infRate: Inflation rate.

--postCode: The postcode of the property's location.

--crimeDensity: The crime density of the property's location.

--crimeRate: The crime rate of the property's location.

--income: The per capita income of the property's location. Calculated as weekly salary.

--distToBusStop: The distance of the property to the nearest bus stop. in kilometers.

--distToCoastline: The distance of the property from the coastline. in kilometers.

--distToNatPark: The distance of the property to the nearest national park. in kilometers.

--distToPark: The distance of the property to the nearest park. in kilometers.

--distToRailLine: The distance of the property from the rail line. in kilometers.

--distToRailStation: The distance of the property from the nearest rail station. in kilometers.

--distToHighway: The distance of the property to the nearest highway. in kilometers.

--distToFreeway: The distance of the property from the nearest freeway. in kilometers.

--distToTunnel: The distance of the property from the undersea tunnel. in kilometers.

--distToMainRoad: The distance of the property from the main road. in kilometers.

--distToSealedRoad: The distance of the property from Sealed Road. in kilometers.

--distToUnsealedRoad: The distance of the property from Unsealed Road. in kilometers.

--airNoise: The air noise of the property's location.

--foreignerRatio: The proportion of foreigners in the location where the property is located.

--distToGPO: The distance of the property from the Sydney GPO. in kilometers.

--NO: Nitric oxide concentration at the location of the property.

--NO2: The concentration of nitrogen dioxide at the location of the property.

--ozone: The ozone concentration at the location of the property.

--neph: Total particulate matter concentration at the location of the property.

--PM10: The concentration of particulate matter less than 10 microns in diameter at the location of the property.

--SO2: The concentration of sulfur dioxide at the location of the property.

--distToAmbulance: The distance of the property from the nearest ambulance station. in kilometers.

--distToFactory: The distance of the property to the nearest factory. in kilometers.

--distToFerry: The distance of the property from the nearest ferry terminal. in kilometers.

--distToHospital: The distance of the property from the nearest hospital. in kilometers.

--distToMedical: The distance of the property to the nearest medical facility. in kilometers.

--distToSchool: The distance of the property to the nearest school. in kilometers.

--distToUniversity: The distance of the property to the nearest university. in kilometers.

2.3 Explore the data

2.3.1 Import the Data

First add a Var.file node to the canvas, and then import SydneyEstateNa.csv into this node.

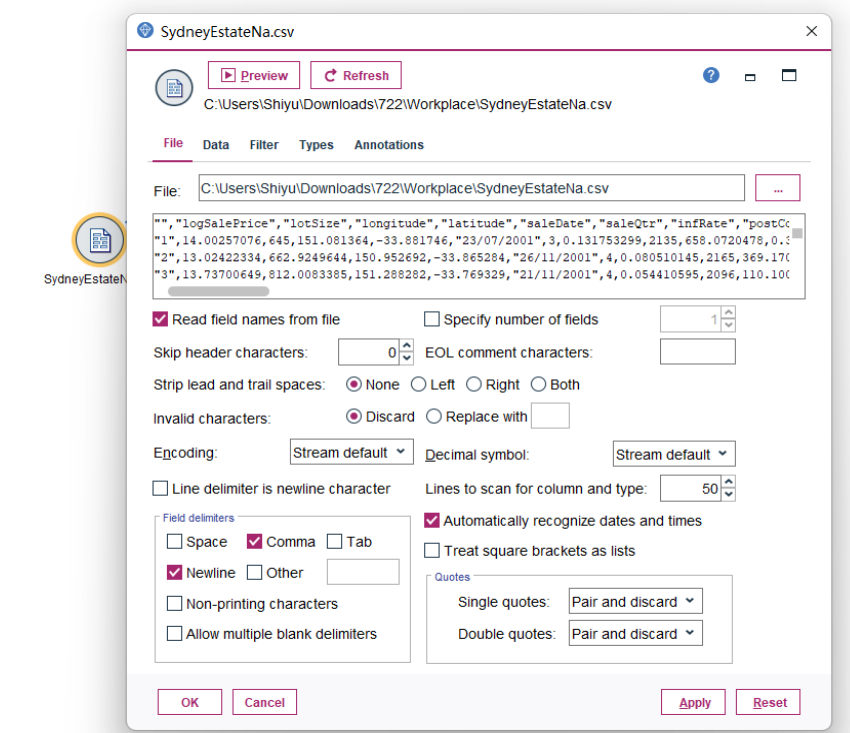


Figure2.3.1: Data Import

2.3.2 Explore and visualize

After importing the data into SPSS Modeler, map the data node to the Data Audit node.

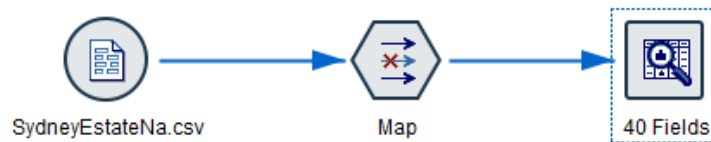


Figure2.3.2: Data Mapping

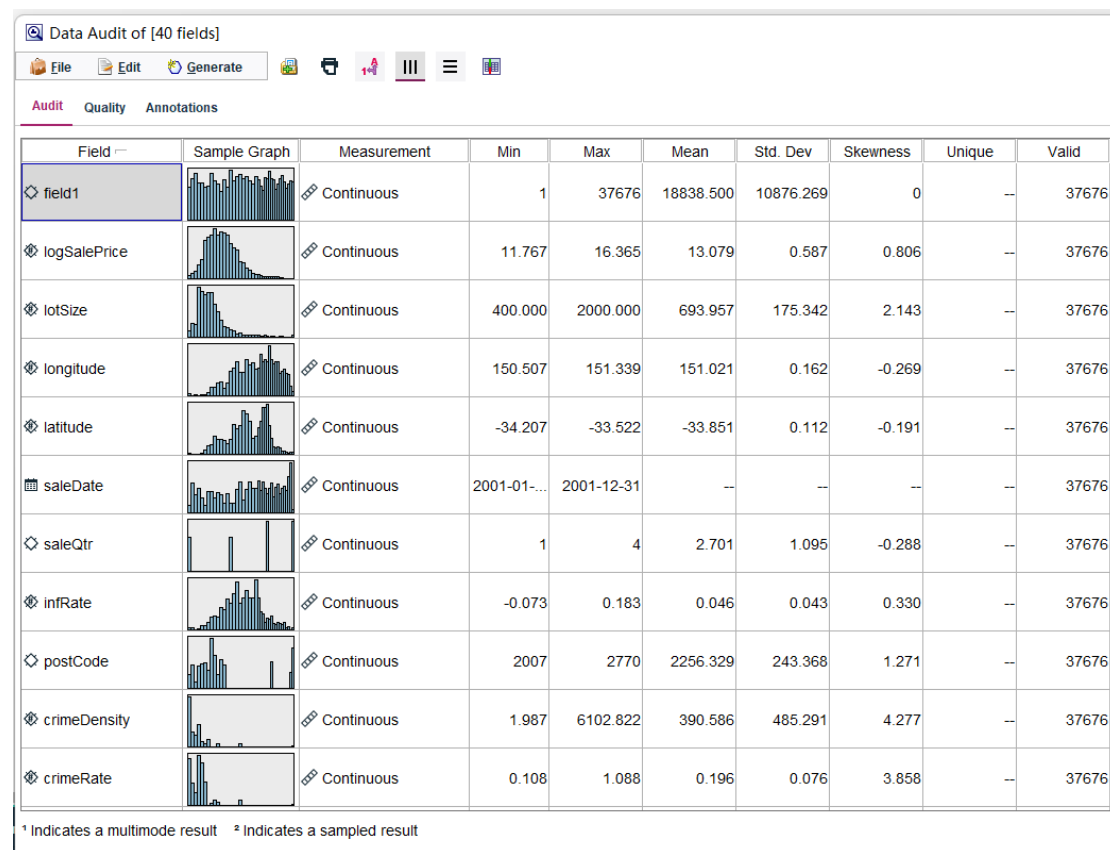


Figure2.3.3: Data Audit

Let's look at the distribution of each variable. The first is the response variable, the house price. Here, house prices are log-transformed.

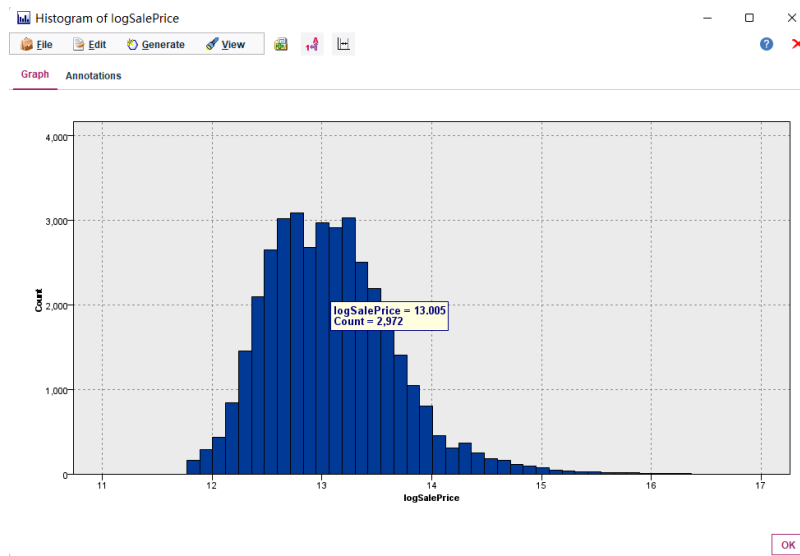


Figure2.3.4: Logged House Price

The housing price is a right-skewed data. Right-skewed data has a long tail that extends to the right. In this case, both the mean and the median are larger than the mode. This is what is expected: everything related to economics is skewed to the right, because there are always some rich people who amass most of the wealth. These outrageous houses are for them.

Next let's look at the lot size.

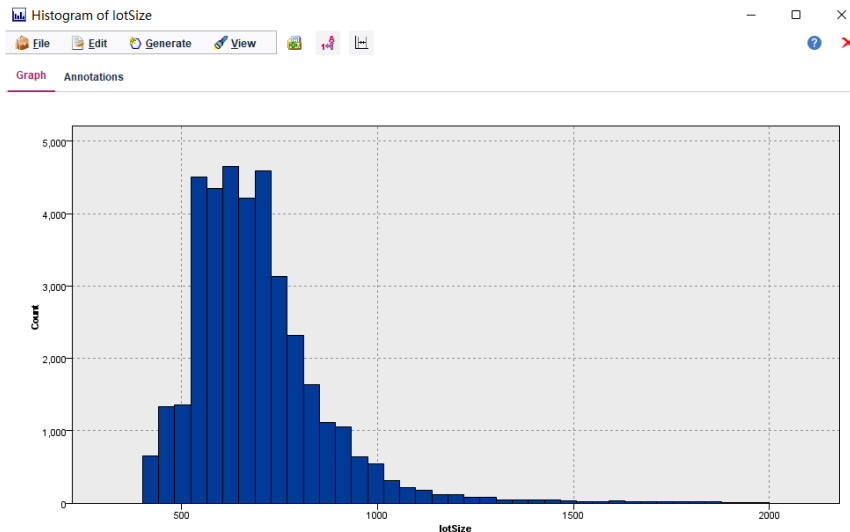


Figure2.3.5: Lot Size

The vast majority of properties in Sydney are between 500 and 1000 square meters. What big houses they are! The largest number of properties are concentrated between 550 square meters and 750 square meters. When calculating the price of a property later, the area is definitely an important point: because we may need to calculate the price of the unit area.

Finally, let's focus on when these properties were sold.

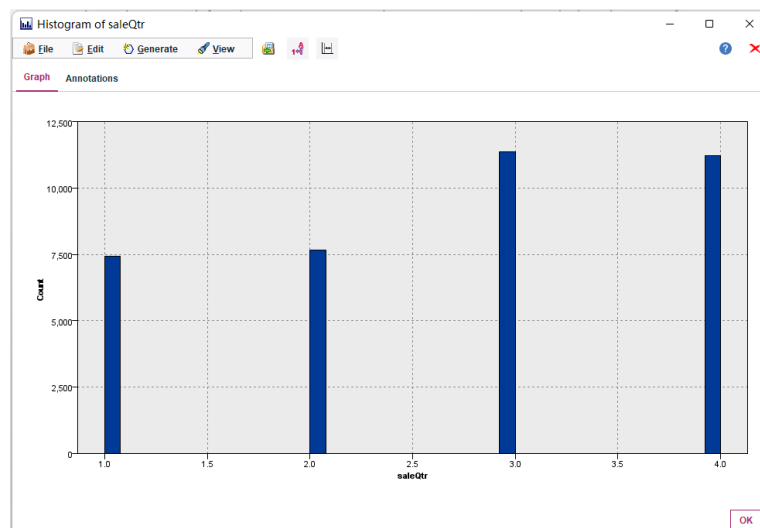


Figure2.3.6: Sale Quarter

The highest number of homes sold in the third quarter, and the least in the first quarter.

2.4 Verify the data quality

The dataset itself is a "clean" dataset. Two of the explanatory variables have missing values, which I added earlier with R.

Data Audit of [40 fields] #5									
Audit Quality Annotations									
Complete fields (%): 95%		Complete records (%): 60.49%							
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value
field1	Continuous	0	0 None	Never	Fixed		100	37676	0
logSalePrice	Continuous	377	15 None	Never	Fixed		100	37676	0
lotSize	Continuous	442	196 None	Never	Fixed		100	37676	0
longitude	Continuous	14	0 None	Never	Fixed		100	37676	0
latitude	Continuous	10	0 None	Never	Fixed		100	37676	0
saleDate	Continuous	0	0 None	Never	Fixed		100	37676	0
saleQtr	Continuous	0	0 None	Never	Fixed		100	37676	0
infRate	Continuous	127	0 None	Never	Fixed		100	37676	0
postCode	Continuous	0	0 None	Never	Fixed		100	37676	0
crimeDensity	Continuous	645	58 None	Never	Fixed		100	37676	0
crimeRate	Continuous	645	58 None	Never	Fixed		100	37676	0
income	Continuous	0	0 None	Never	Fixed		100	37676	0
distToBusStop	Continuous	293	59 None	Never	Fixed		80.075	30169	7507
distToCoastline	Continuous	211	1 None	Never	Fixed		74.013	27885	9791
distToNatPark	Continuous	476	0 None	Never	Fixed		100	37676	0
distToPark	Continuous	376	355 None	Never	Fixed		100	37676	0
distToRailLine	Continuous	1168	1 None	Never	Fixed		100	37676	0
distToRailSta	Continuous	1144	0 None	Never	Fixed		100	37676	0
distToHighway	Continuous	735	6 None	Never	Fixed		100	37676	0
distToFreeway	Continuous	641	87 None	Never	Fixed		100	37676	0
distToTunnel	Continuous	41	0 None	Never	Fixed		100	37676	0
distToMainR	Continuous	532	46 None	Never	Fixed		100	37676	0
distToSealed	Continuous	415	55 None	Never	Fixed		100	37676	0
distToUnseal	Continuous	31	0 None	Never	Fixed		100	37676	0
airNoise	Continuous	0	1066 None	Never	Fixed		100	37676	0
foreignerRatio	Continuous	190	0 None	Never	Fixed		100	37676	0
distToGPO	Continuous	49	0 None	Never	Fixed		100	37676	0
NO	Continuous	2453	0 None	Never	Fixed		100	37676	0
NO2	Continuous	2453	0 None	Never	Fixed		100	37676	0
ozone	Continuous	0	21 None	Never	Fixed		100	37676	0
neph	Continuous	0	0 None	Never	Fixed		100	37676	0
PM10	Continuous	0	0 None	Never	Fixed		100	37676	0

Figure2.4.1: Data Quality

All variable types are Continuous. This is because the variables in the dataset are all described numerically. There are some that should be considered categorical, though, such as saleDate: the quarter of the property sold, as it has only four possible types of values.

There are two variables with incomplete data (Null Value) that we need to deal with.

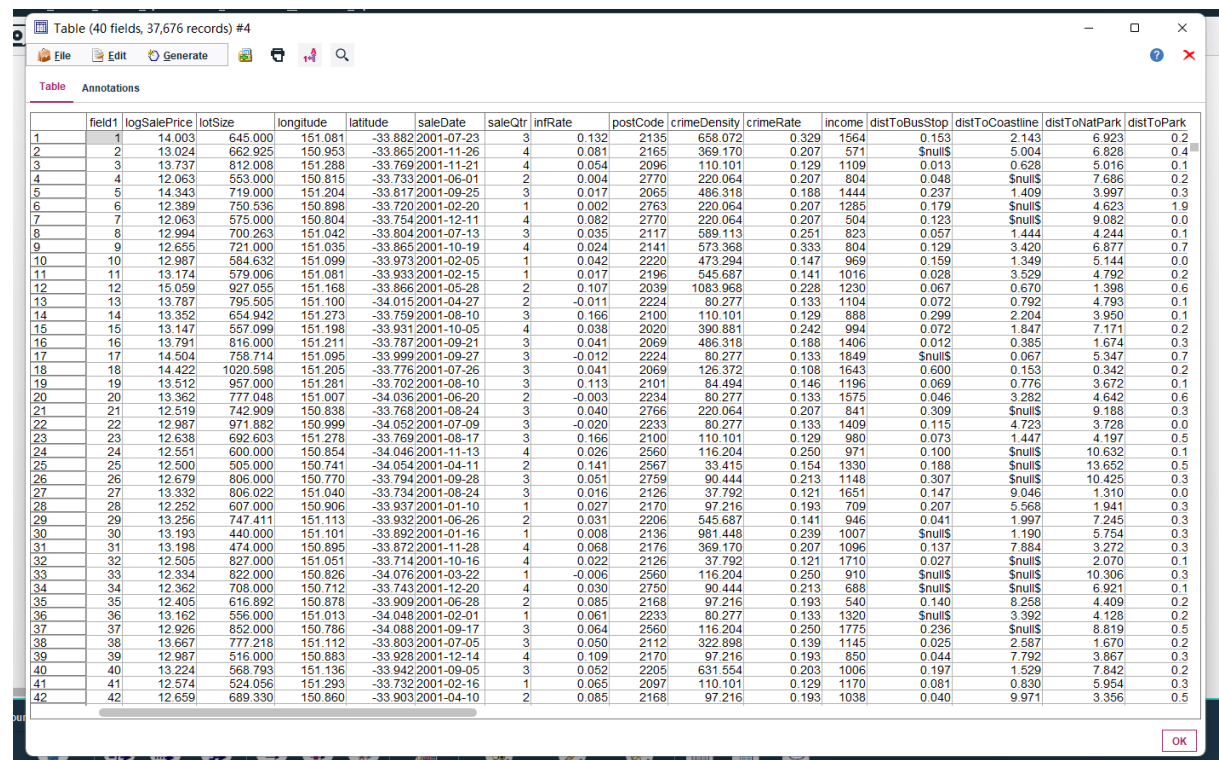
For the distToBusStop variable, I randomly lost 20% of the data; for distToCoastline, I used the loss probability of $\text{logit}(p) = \log(p/(1-p)) = -10 + y$, where y refers to the value of distToCoastline value. So distToBusStop is Missing At Random (MAR) , while distToCoastline is Missing Not At Random (MNAR).

In addition, there are some outliers and extreme values detected in the data. In a later analysis, I will deduce whether they are really outliers or extreme values.

3. Data preparation

3.1 Select the data

By linking a Table node to the original data, we can see that our original data now looks like this:



	field1	logSalePrice	lotSize	longitude	latitude	saleDate	saleQtr	infRate	postCode	crimeDensity	crimeRate	income	distToBusStop	distToCoastline	distToNatPark	distToPark
1	1	14.003	645.000	151.081	-33.882	2001-07-23	3	0.132	2135	658.072	0.329	1564	0.153	2.143	6.923	0.2
2	2	13.024	662.925	150.953	-33.865	2001-11-26	4	0.081	2165	369.170	0.207	571	\$null\$	5.004	6.828	0.4
3	3	13.737	812.008	151.288	-33.769	2001-11-21	4	0.054	2096	110.101	0.129	1109	0.013	0.628	5.016	0.1
4	4	12.063	553.000	150.815	-33.733	2001-06-01	2	0.004	2770	220.064	0.207	804	0.048	\$null\$	7.686	0.2
5	5	14.343	719.000	151.204	-33.817	2001-09-25	3	0.017	2065	486.318	0.188	1444	0.237	1.409	3.997	0.3
6	6	12.389	750.536	150.898	-33.720	2001-02-20	1	0.002	2763	220.064	0.207	1285	0.179	\$null\$	4.623	1.9
7	7	12.063	575.000	150.804	-33.754	2001-12-11	4	0.082	2770	220.064	0.207	504	0.123	\$null\$	9.082	0.0
8	8	12.994	700.263	151.042	-33.804	2001-07-13	3	0.035	2117	589.113	0.251	823	0.057	1.444	4.244	0.1
9	9	12.655	721.000	151.035	-33.865	2001-10-19	4	0.024	2141	573.368	0.333	804	0.129	3.420	6.877	0.7
10	10	12.987	584.632	151.099	-33.973	2001-02-05	1	0.042	2220	473.294	0.147	969	0.159	1.349	5.144	0.0
11	11	13.174	579.006	151.081	-33.933	2001-02-15	1	0.017	2196	545.687	0.141	1016	0.028	3.529	4.792	0.2
12	12	15.059	927.055	151.168	-33.866	2001-05-28	2	0.107	2039	1083.968	0.228	1230	0.067	0.670	1.398	0.6
13	13	13.787	795.505	151.100	-34.015	2001-04-27	2	-0.011	2224	80.277	0.133	1104	0.072	0.792	4.793	0.1
14	14	13.352	654.942	151.273	-33.759	2001-08-10	3	0.166	2100	110.101	0.129	888	0.299	2.204	3.950	0.1
15	15	13.147	557.099	151.198	-33.931	2001-10-05	4	0.038	2020	390.881	0.242	994	0.072	1.847	7.171	0.2
16	16	13.791	816.000	151.211	-33.787	2001-09-21	3	0.041	2069	486.318	0.188	1406	0.012	0.385	1.674	0.3
17	17	14.504	758.714	151.095	-33.999	2001-09-27	3	-0.012	2224	80.277	0.133	1849	\$null\$	0.067	5.347	0.7
18	18	14.422	1020.598	151.205	-33.776	2001-07-26	3	0.041	2069	126.372	0.108	1643	0.600	0.153	0.342	0.2
19	19	13.512	957.000	151.281	-33.702	2001-08-10	3	0.113	2101	84.494	0.146	1196	0.069	0.776	3.672	0.1
20	20	13.362	777.048	151.007	-34.036	2001-06-20	2	-0.003	2234	80.277	0.133	1575	0.046	3.282	4.642	0.6
21	21	12.519	742.909	150.838	-33.768	2001-08-24	3	0.040	2766	220.064	0.207	841	0.309	\$null\$	9.188	0.3
22	22	12.987	971.882	150.999	-34.052	2001-07-09	3	-0.020	2233	80.277	0.133	1409	0.115	4.723	3.728	0.0
23	23	12.638	692.603	151.278	-33.769	2001-08-17	3	0.166	2100	110.101	0.129	980	0.073	1.447	4.197	0.5
24	24	12.551	600.000	150.854	-34.046	2001-11-13	4	0.026	2560	116.204	0.250	971	0.100	\$null\$	10.632	0.1
25	25	12.500	505.000	150.741	-34.054	2001-04-11	2	0.141	2567	33.415	0.154	1330	0.188	\$null\$	13.652	0.5
26	26	12.679	806.000	150.770	-33.794	2001-09-28	3	0.051	2759	90.444	0.213	1148	0.307	\$null\$	10.425	0.3
27	27	13.332	806.022	151.040	-33.734	2001-08-24	3	0.016	2126	37.792	0.121	1651	0.147	9.046	1.310	0.0
28	28	12.252	607.000	150.906	-33.937	2001-01-10	1	0.027	2170	97.216	0.193	709	0.207	5.568	1.941	0.3
29	29	13.256	747.411	151.113	-33.932	2001-06-26	2	0.031	2206	545.687	0.141	946	0.041	1.997	7.245	0.3
30	30	13.193	440.000	151.101	-33.892	2001-01-16	1	0.008	2136	981.448	0.239	1007	\$null\$	1.190	5.754	0.3
31	31	13.198	474.000	150.895	-33.872	2001-11-28	4	0.068	2176	369.170	0.207	1096	0.137	7.884	3.272	0.3
32	32	12.505	827.000	151.051	-33.714	2001-10-16	4	0.022	2126	37.792	0.121	1710	0.027	\$null\$	2.070	0.1
33	33	12.334	622.000	150.826	-34.076	2001-03-22	1	-0.006	2560	116.204	0.250	910	\$null\$	\$null\$	10.306	0.3
34	34	12.952	708.000	150.712	-33.743	2001-12-20	4	0.030	2750	90.444	0.213	688	\$null\$	\$null\$	6.921	0.1
35	35	12.405	616.892	150.878	-33.909	2001-06-28	2	0.085	2168	97.216	0.193	540	0.140	8.258	4.409	0.2
36	36	13.162	556.000	151.013	-34.048	2001-02-01	1	0.061	2233	80.277	0.133	1320	\$null\$	3.392	4.128	0.2
37	37	12.926	852.000	150.786	-34.088	2001-09-17	3	0.064	2560	116.204	0.250	1775	0.236	\$null\$	8.819	0.5
38	38	13.667	777.218	151.112	-33.803	2001-07-05	3	0.050	2112	322.898	0.139	1145	0.025	2.587	1.670	0.2
39	39	12.987	516.000	150.883	-33.928	2001-12-14	4	0.109	2170	97.216	0.193	850	0.044	7.792	3.867	0.3
40	40	13.224	568.793	151.136	-33.942	2001-09-05	3	0.052	2205	631.554	0.203	1006	0.197	1.529	7.842	0.2
41	41	12.574	524.056	151.293	-33.732	2001-02-16	1	0.065	2097	110.101	0.129	1170	0.081	0.830	5.954	0.3
42	42	12.659	689.330	150.860	-33.903	2001-04-10	2	0.085	2168	97.216	0.193	1038	0.040	9.971	3.356	0.5

Figure3.1.1: Raw data overview

Select rows. Each row of data contains a real estate transaction information. Our study is all Sydney property transactions, and more data leads to a better fit, so I decided to keep data for all rows.

Select columns. All the columns here are potentially valuable variables that may be helpful in predicting real estate prices. Since the data does not contain sensitive customer information, all columns can be preserved. The one exception is the first column "field1", which is the ordinal of the rows and isn't helpful for predicting the real estate prices we care about.

In addition, there are some variables that are more suitable to be treated as categorical variables, such as fiscal quarter, zip code, and transaction date. However, when used as categorical variables, postal code and transaction date will have too many levels, which is not conducive to us building a stable model, and will also bring trouble to interpret the statistical significance of the model, so I choose to discard them.

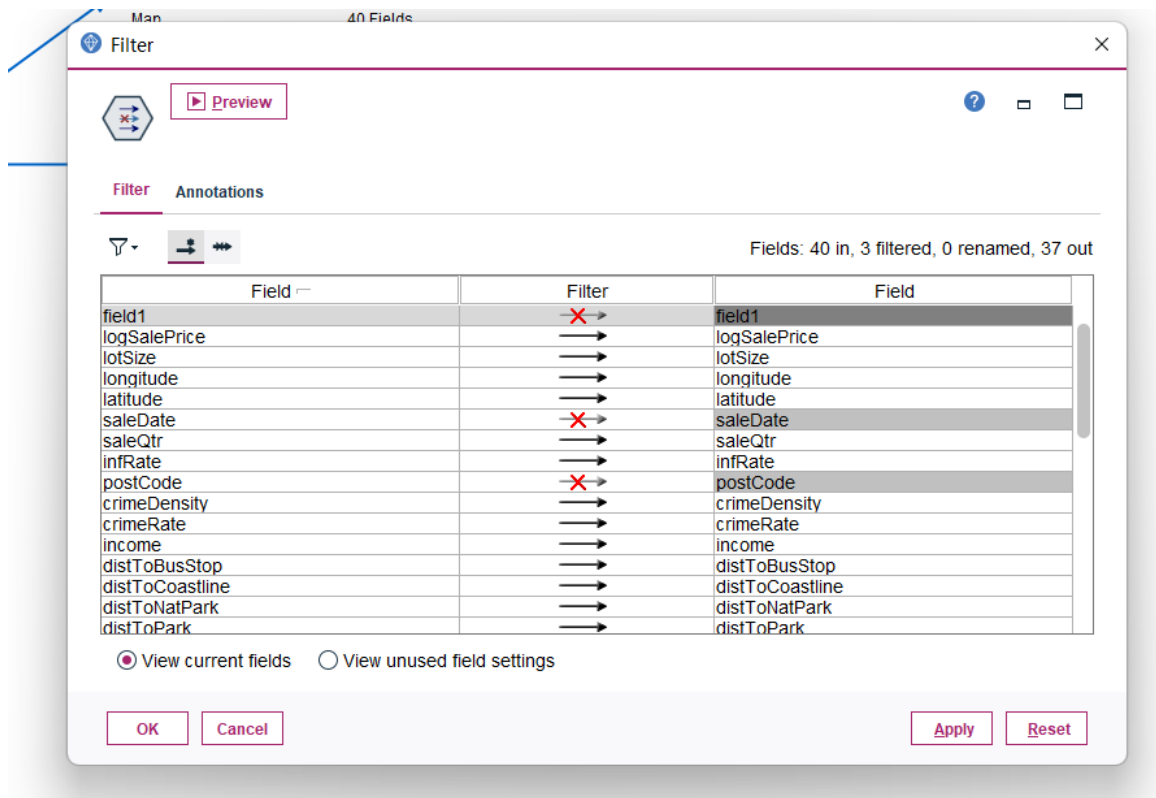


Figure3.1.2: Column Filter

3.2 Clean the data

As previously described, only two columns in the dataset contain missing data. Unfortunately, based on my intuition, they are in two columns that are important for predicting property prices, and simply dropping those two columns is not a good idea.

We could delete all rows with data, but I'd like to try to avoid doing that. Doing so will lose a lot of data, potentially making our model less robust; one of the columns is missing non-randomly (MNAR), and deleting the corresponding row may lead to bias in our final model predictions.[5]

Here I use a Data Audit Node again to check the current data:

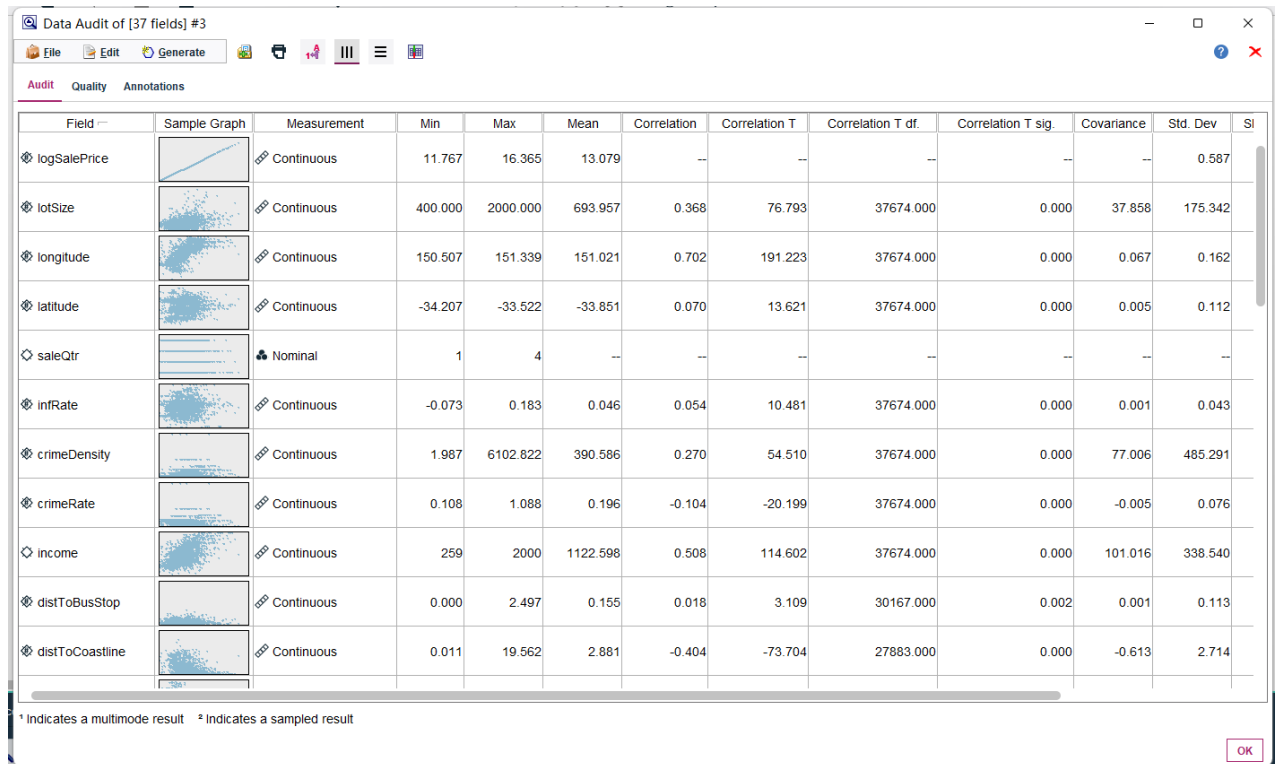


Figure3.2.1: Data Audit results

Now, we need to apply an appropriate imputation method for missing values. For the two columns with missing values, I decided to use the default algorithm to Impute. As for outliers and extremes, I coerce outliers and discard extreme points.

Generate a SuperNode to execute the algorithm:

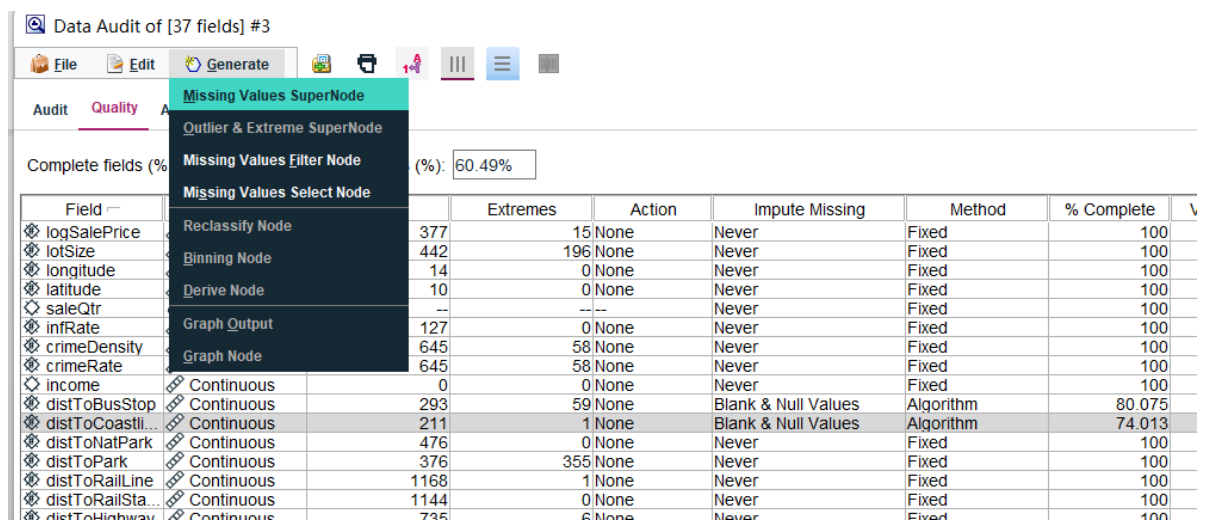


Figure3.2.2: Data Imputation

Run Data Audit again, all missing values have been filled. Extreme points no longer exist, and outliers are reduced.

Data Audit of [37 fields] #17

File Edit Generate

Audit **Quality** Annotations

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid F
logSalePrice	Continuous	398	0 None		Never	Fixed	100	
lotSize	Continuous	623	0 None		Never	Fixed	100	
longitude	Continuous	0	0 None		Never	Fixed	100	
latitude	Continuous	0	0 None		Never	Fixed	100	
saleQtr	Nominal	--	--		Never	Fixed	100	
infRate	Continuous	0	0 None		Never	Fixed	100	
crimeDensity	Continuous	976	0 None		Never	Fixed	100	
crimeRate	Continuous	643	0 None		Never	Fixed	100	
income	Continuous	0	0 None		Never	Fixed	100	
distToBusStop	Continuous	605	0 None		Never	Fixed	100	
distToCoastli...	Continuous	0	0 None		Never	Fixed	100	
distToNatPark	Continuous	0	0 None		Never	Fixed	100	
distToPark	Continuous	786	0 None		Never	Fixed	100	
distToRailLine	Continuous	1114	0 None		Never	Fixed	100	
distToRailSta...	Continuous	1068	0 None		Never	Fixed	100	
distToHighway	Continuous	1105	0 None		Never	Fixed	100	
distToFreeway	Continuous	888	0 None		Never	Fixed	100	
distToTunnel	Continuous	0	0 None		Never	Fixed	100	
distToMainR...	Continuous	558	0 None		Never	Fixed	100	
distToSealed...	Continuous	516	0 None		Never	Fixed	100	
distToUnseal...	Continuous	0	0 None		Never	Fixed	100	
airNoise	Continuous	0	0 None		Never	Fixed	100	
foreignerRatio	Continuous	168	0 None		Never	Fixed	100	
distToGPO	Continuous	0	0 None		Never	Fixed	100	
NO	Continuous	2360	0 None		Never	Fixed	100	
NO2	Continuous	2360	0 None		Never	Fixed	100	
ozone	Continuous	0	0 None		Never	Fixed	100	
neph	Continuous	0	0 None		Never	Fixed	100	
PM10	Continuous	0	0 None		Never	Fixed	100	
SO2	Continuous	0	0 None		Never	Fixed	100	
distToAmbul...	Continuous	1098	0 None		Never	Fixed	100	
distToFactory	Continuous	73	0 None		Never	Fixed	100	
distToFerry	Continuous	193	0 None		Never	Fixed	100	
distToHospital	Continuous	604	0 None		Never	Fixed	100	
distToMedical	Continuous	746	0 None		Never	Fixed	100	
distToSchool	Continuous	394	0 None		Never	Fixed	100	
distToLake	Continuous	609	0 None		Never	Fixed	100	

Figure3.2.3: Data Imputation results

3.3 Construct the data

The real estate prices in the original data are on the log-scale. This can be useful when modeling, but is less intuitive when interpreting. So, I'm adding a new variable here: property price (thousands of AUD). This variable is obtained by taking the exponent of "logSalePrice" and dividing it by 1000.

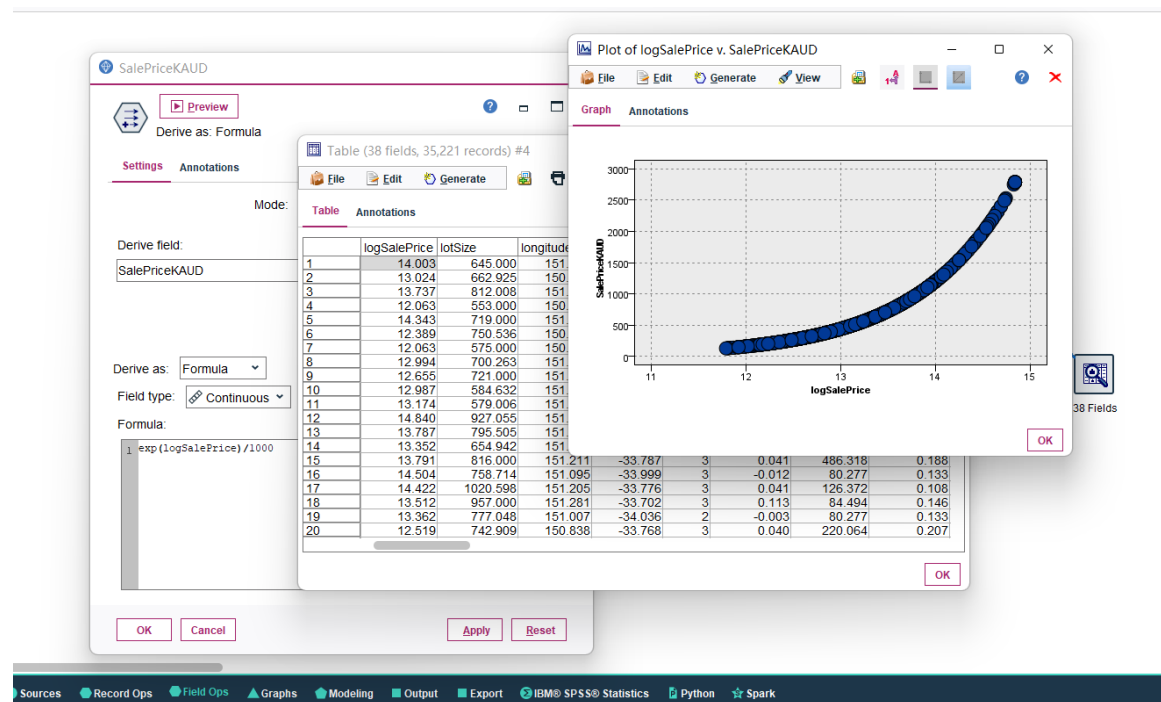


Figure3.3.1: Generate new column

It can be seen from the plot that the restored real estate prices show an exponential relation to log-Sale Price.

3.4 Integrate various data sources

Our data comes from a single CSV file, and the newly added variables have already been integrated into the current data frame, so here I have to break the data into halves and merge it back. Since here I am splitting on the rows (rather than columns), the Append node should be used to merge the data.

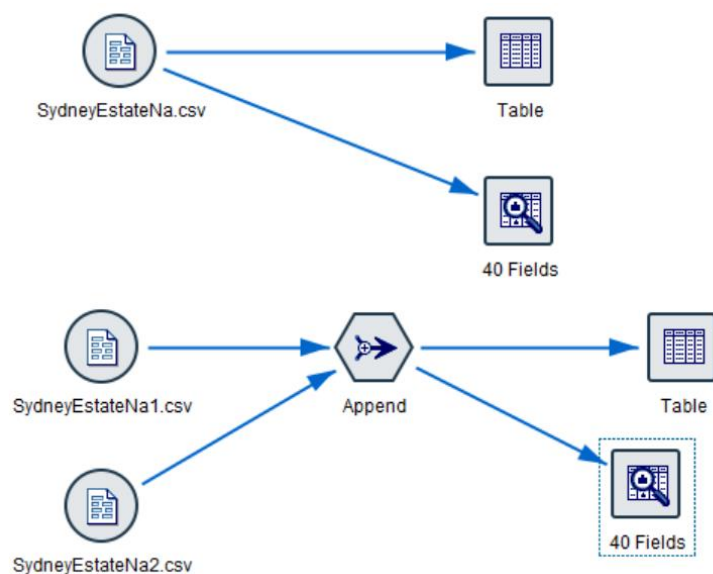


Figure3.4.1: Merge the Data

The merged dataset is the same as the original dataset.

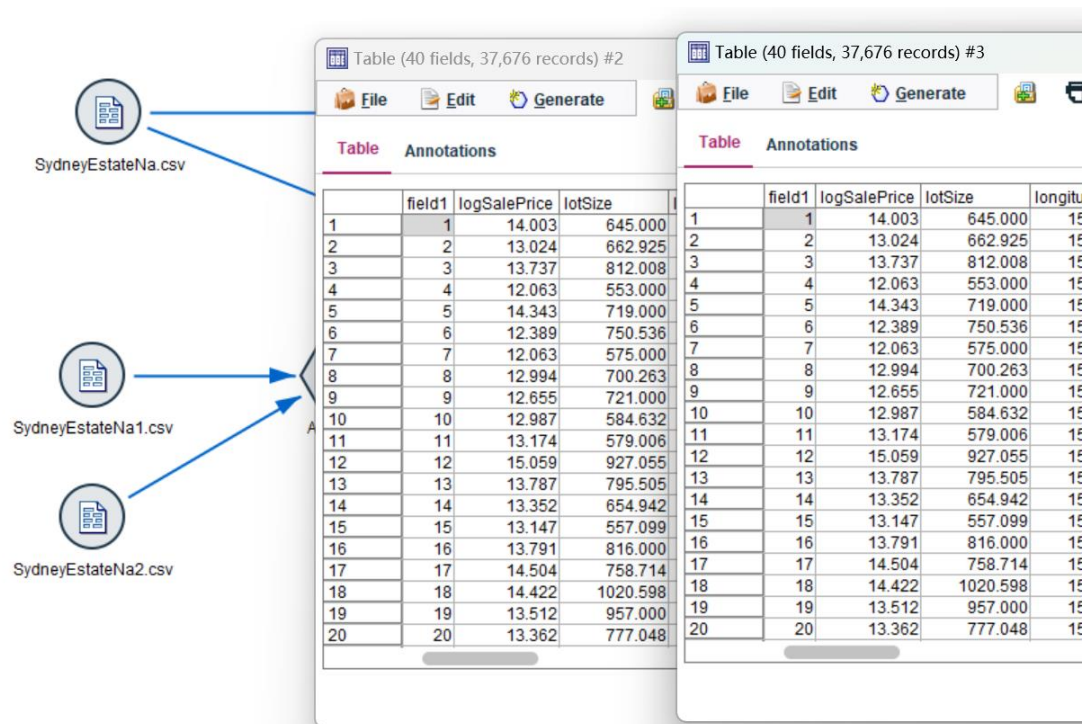


Figure3.4.2: Merged Data

3.5 Format the data as required

Financial quarter is more suitable to be considered a nominal variable because it is discrete and has only 4 levels.

In addition, input and target should also be specified. input refers to the explanatory variable and target refers to the response variable.

I used a Type Node to handle the filtered data.

Type

Preview

Types Format Annotations

Read Values Clear Values Clear All Values

Field	Measurement	Values	Missing	Check	Role
logSalePrice	Continuous	[11.76746515,16...		None	Target
lotSize	Continuous	[400.0,2000.0]		None	Input
longitude	Continuous	[150.507156,151...		None	Input
latitude	Continuous	[-34.207357,-33.5...		None	Input
saleQtr	Nominal	1,2,3,4		None	Input
infRate	Continuous	[-0.073005101,0.1...		None	Input
crimeDensity	Continuous	[1.986585638,610...		None	Input
crimeRate	Continuous	[0.107858013,1.0...		None	Input
income	Continuous	[259,2000]		None	Input
distToBusStop	Continuous	[3.81042E-4,2.49...		None	Input
distToCoastline	Continuous	[0.010774533,19...		None	Input
distToNatPark	Continuous	[0.04300466,14.9...		None	Input
distToPark	Continuous	[0.00406563,5.98...		None	Input
distToRailLine	Continuous	[0.016473452,15...		None	Input
distToRailStation	Continuous	[0.02927692,16...		None	Input

Figure3.5.1: Type Node

Then, I used a Map Node to rename the columns. I thought the previous "SalePriceKAUD" was verbose, here I changed it to "SalePrice".

Map

Preview

Filter Annotations

Field	Filter
distToTunnel	distToTunnel
distToMainRoad	distToMainRoad
distToSealedRoad	distToSealedRoad
distToUnsealedRoad	distToUnsealedRoad
airNoise	airNoise
foreignerRatio	foreignerRatio
distToGPO	distToGPO
NO	NO
NO2	NO2
ozone	ozone
neph	neph
PM10	PM10
SO2	SO2
distToAmbulance	distToAmbulance
distToFactory	distToFactory
distToFerry	distToFerry
distToHospital	distToHospital
distToMedical	distToMedical
distToSchool	distToSchool
distToUniversity	distToUniversity
SalePriceKAUD	SalePrice

View current fields View unused field settings

OK Cancel

Data Audit of [38 fields] #2

File Edit Generate

Audit Quality Annotations

Field	Sample Graph	Measurement	Min	Max
distToAmbulance		Continuous	0.034	
distToFactory		Continuous	0.010	
distToFerry		Continuous	0.070	
distToHospital		Continuous	0.028	
distToMedical		Continuous	0.007	
distToSchool		Continuous	0.009	
distToUniversity		Continuous	0.025	
SalePrice		Continuous	128.987	27

* Indicates a multimode result * Indicates a sampled result

Figure3.5.2: Rename new column

4. Data transformation

4.1 Reduce the data

Feature selection helps identify the most important fields when modeling and predicting outcomes. This is useful when there are many explanatory variables, because not every explanatory variable will be helpful for modeling and prediction, and useless explanatory variables will increase the computational cost and affect the stability of the model.

Here, I added a Feature Selection Node and performed feature selection according to the default configuration. Here is the result:

The screenshot shows the 'SalePrice' dialog box with the 'Model' tab selected. The 'Rank' dropdown is set to 'Rank'. The table below shows the results of the feature selection process.

Rank	Field	Measurement	Importance	Value
13	crimeDensity	Continuous	Important	1.0
14	distToHospital	Continuous	Important	1.0
15	distToFactory	Continuous	Important	1.0
16	distToRailLine	Continuous	Important	1.0
17	distToRailStation	Continuous	Important	1.0
18	ozone	Continuous	Important	1.0
19	infRate	Continuous	Important	1.0
21	neph	Continuous	Important	1.0
22	distToMainRoad	Continuous	Important	1.0
23	distToAmbulance	Continuous	Important	1.0
24	distToMedical	Continuous	Important	1.0
25	distToPark	Continuous	Important	1.0
26	crimeRate	Continuous	Important	1.0
27	distToUniversity	Continuous	Important	1.0
28	distToSchool	Continuous	Important	1.0
29	distToHighway	Continuous	Important	1.0
30	distToFreeway	Continuous	Important	1.0
31	distToSealedRoad	Continuous	Important	1.0
32	foreignerRatio	Continuous	Important	0.999
33	distToBusStop	Continuous	Unimportant	0.642

Selected fields: 31 Total fields available: 36

Thresholds: ☒ > 0.95 ☒ <= 0.95 ☒ < 0.9

4 Screened Fields

Field	Measurement	Reason
PM10	Continuous	Coefficient of variation below threshold
longitude	Continuous	Coefficient of variation below threshold
latitude	Continuous	Coefficient of variation below threshold
airNoise	Continuous	Coefficient of variation below threshold

Buttons: OK, Cancel, Apply, Reset

Figure4.1.1: Feature Selection

Four variables were discarded, and an additional variable was considered "unimportant".

Intuitively, the distance from the bus stop should be an important variable, but since it is non-randomly lost, the previous algorithm did not fill in the missing values well, which may be the reason why it is judged to be unimportant here.

We follow the feature selection node's suggestion, keeping explanatory variables deemed "important". There are 31 important explanatory variables.

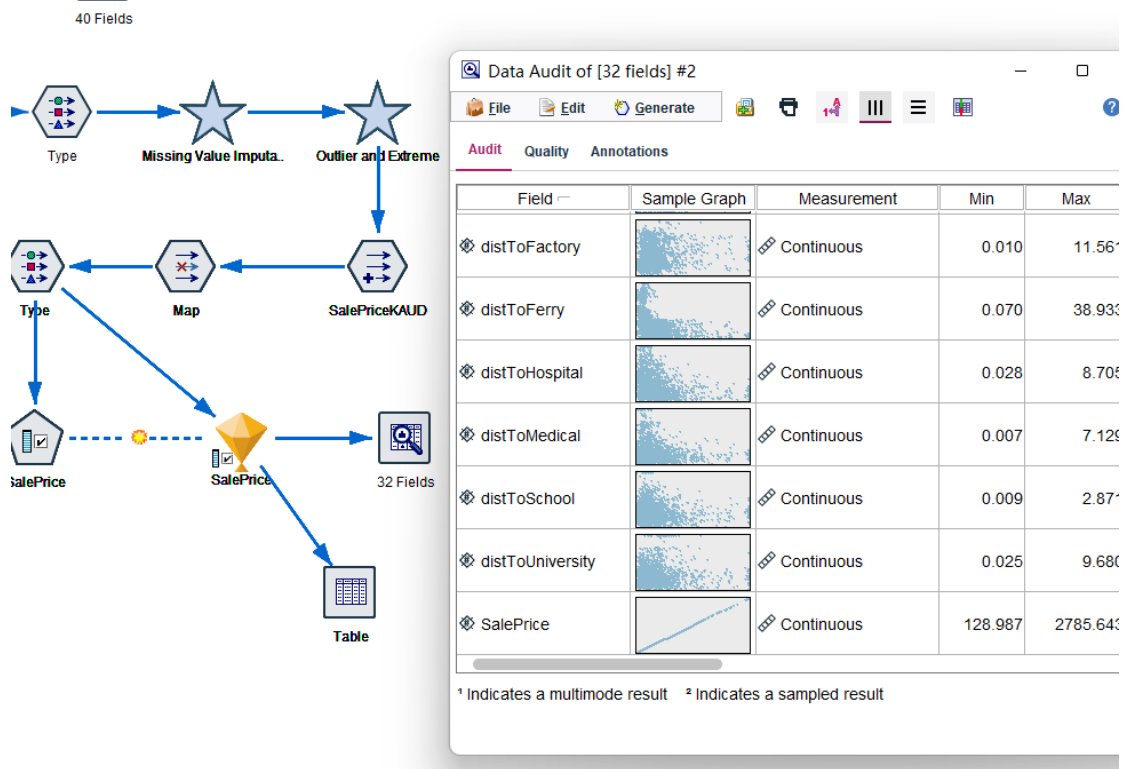


Figure4.1.2: Feature Selection results

4.2 Project the data

From the previous results, the relationship between property prices and most of the explanatory variables is the inverse of the log (and its multiples). Therefore, it would be a better idea to use log-scale explanatory variables.

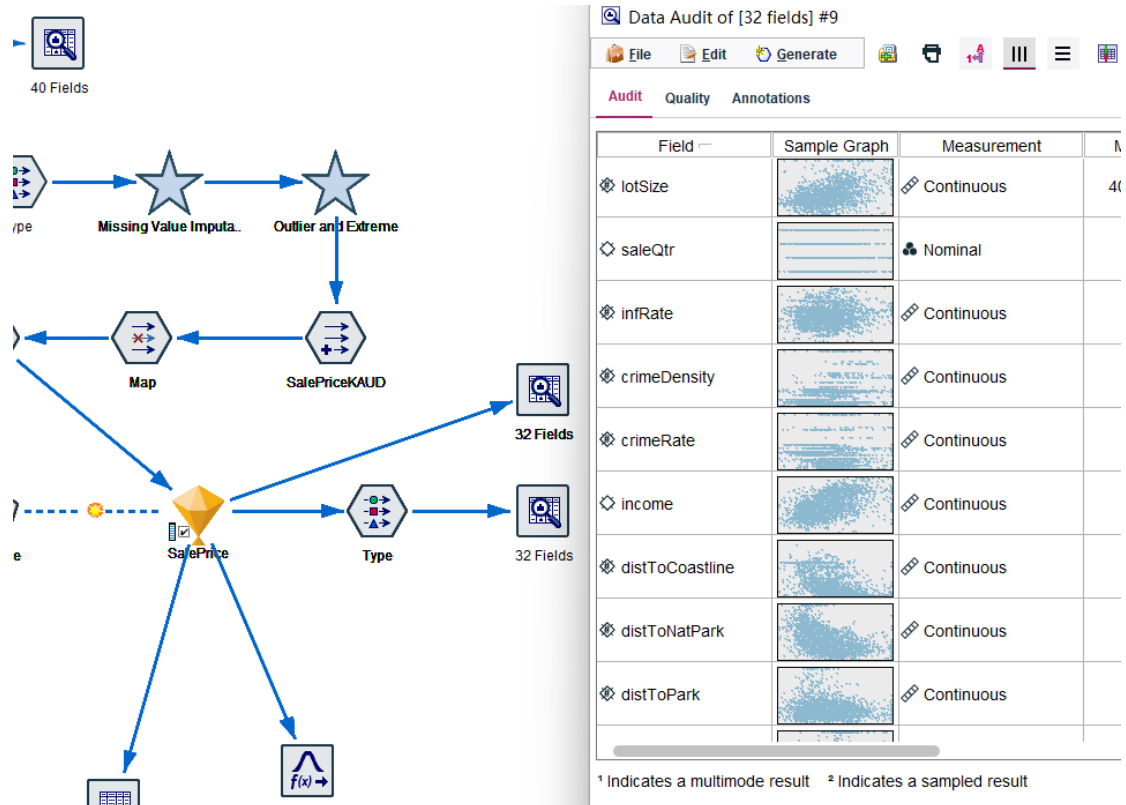


Figure 4.2.1: Relationship between log-price and explanatory variables

Log-transformed data generally better fit the model assumptions. Reciprocal-shaped distributions are difficult to interpret, whereas linear-like distributions are easier to fit and interpret.[6]

From the above figure, the relationship between explanatory variables and response variables is clearer: for example, income and house prices have a linear positive correlation.

5. Data-mining algorithm(s) selection

5.1 Match and discuss the objectives of data mining to data mining methods

The goal of this research project is to study the factors that influence Sydney property prices and to predict property prices based on different explanatory variables in the dataset.

The data mining goals we focus on are:

Explanation: Identify important explanatory variables that affect property prices and explain the relationship between these variables and property prices.

Prediction: Predict the house price given the explanatory variables.

Desired Outcome: Get to know what factors affect house prices the most, and what's trivial, and make the model obtain a high prediction accuracy in the test set.

There are three main categories of data mining methods: **Classification**, **Clustering** and **Regression**. [7] we can implement all these algorithms in SPSS Modeler. The three types of data mining methods will be explained below and combined with our data mining goals to see why only regression is the appropriate data mining method for this project.

Classification is a supervised learning method. Classification models are suitable for datasets where the response variable is discrete. [7]

Clustering is an unsupervised learning method. Unsupervised learning is a training method of machine learning, which is essentially a statistical method, a training method in which some potential structures can be found in unlabeled data. [7] Clustering focuses on identifying similar data and labeling the data according to the group to which the data corresponds.

Regression is a common data analysis method. Regression models produce continuous response variables. [7] Common regression models include linear regression, generalized linear regression, random forest regression, and lasso regression.

5.2 Select the appropriate data-mining method(s) based on discussion

In this project, we will use a variety of models and data mining methods and come up with the most suitable model or models for prediction.

Based on the objectives of data mining,

1. For **classification**, the response variable we want to model and predict is house price (or its logarithm), which is a continuous variable rather than a categorical variable, and if it

were a categorical variable, there would be close to the number of rows of data levels, thus making it impossible to interpret the modeling results and make predictions. Therefore, **classification is not a method that meets our data mining goals.**

2. **Clustering** is an unsupervised learning method with no predefined outputs, is suitable when the original dataset does not contain a response variable. We were also unable to verify the predictive accuracy of the clustering, as there was no established criterion. This does not meet our data mining goals, so **clustering is not the data mining method we should choose here.**
3. For **regression**, the response variable here is house price (or its logarithm), which is a continuous variable, regression methods are suitable in this case. So, **I decided to use regression as the data mining method we will use.**

6. Data-mining algorithm(s) selection

6.1 Conduct exploratory analysis and discuss

As we discussed earlier, we will use regression as the data mining method. There are many data mining methods for regression, such as regression trees, random forests, Bayesian regression, etc.

Regression trees are basically decision trees for regression tasks that can be used to predict continuous valued outputs instead of discrete outputs.[12] The basic idea behind the algorithm is to find the point in the independent variable, split the dataset into 2 parts so that the mean squared error is minimum at that point. The algorithm does this in an iterative manner, forming a tree-like structure.[12]

Random forests grow many classification trees. To classify new objects in the input vector, place the input vector under each tree in the forest. Each tree is given a classification, and we say the tree "votes" for that class. The forest selects the class with the most votes (amongst all trees in the forest).[13]

There are two types of machine learning: supervised learning and unsupervised learning.

Supervised learning is a machine learning method defined by the use of labeled datasets. These datasets are designed to train or "supervised" algorithms to classify data or accurately predict outcomes. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.[14]

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns in data without human intervention (thus, they are "unsupervised").[14]

The goal of this research is to study the factors that affect the price of real estate, build a model and use these factors to predict the price of real estate, so the input of the model is some explanatory variables, and the output is the house price. This is a scenario with a clear output target, so we should choose a supervised model.

6.2 Select data-mining algorithms based on discussion

Specifically, there are some models that are more traditional and straightforward, such as multivariate regression and linear models, which are helpful for our later explanations. There are models that use more modern statistical theory, which may perform better, but also incur greater computational overhead. Here, we will choose some models and proceed to the subsequent modeling steps.

In fact, all models available in SPSS Modeler are our potential models. This includes Regression, Generalized Linear, KNN, SVM, etc. We can fit all the models and compare their predictive

performance on the input data, but fitting all the models is computationally expensive.

I decide to fit three models: **Linear model**, **Multivariate Regression model** and **Random Forest**.

Linear regression is the most traditional one of regression methods, which assumes a linear relationship between the explanatory variable and the response variable. Here, I will choose a single variable as the explanatory variable.

Multivariate Regression can tell the relationship between multiple explanatory variables and a single response variable. Here, I will put all explanatory variables (after the feature selection) into the model.

Random forest belongs to the Bagging algorithm. The general idea is to train multiple weak models and package them to form a strong model. The performance of the strong model is much better than that of a single weak model.[8] In the training phase, random forest uses bootstrap sampling to collect multiple different sub-training datasets from the input training dataset to train multiple different decision trees in turn; in the prediction phase, random forest averages the prediction results of multiple internal decision trees to obtain final result.[8]

6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

6.3.1 Linear regression model

For a linear model, the response variable is house prices (thousands of AUD) and the explanatory variable will be one of the variables in the dataset. Intuitively, I think that lot size will have a linear positive relationship with house prices: the bigger the house, the more expensive it will be.

For a linear model, the response variable is house prices (thousands of AUD) and the explanatory variable will be one of the variables in the dataset. Intuitively, I think that lot size will have a linear positive relationship with house prices: the bigger the house, the more expensive it will be.

I used a regression node to represent this model. The explanatory variable is lot size, while the response variable is house prices (thousands of AUD).

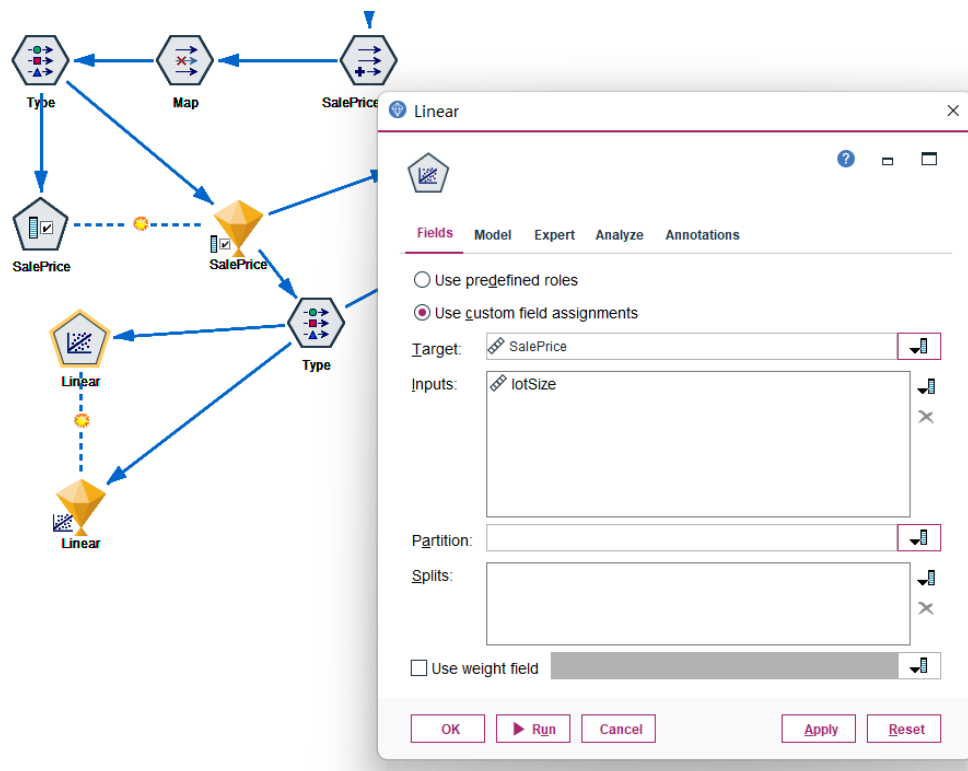


Figure6.3.1: Linear Model Parameter Settings

6.3.2 Multivariate Regression model

For the Multivariate Regression model, I used all the explanatory variables previously filtered as input, and the output (response) variable being house price (AUD) on log scale. This is obviously a more complex model, with a high probability that it will perform better than the linear model.

To include all explanatory variables in the input data, simply keep the regression node preset.

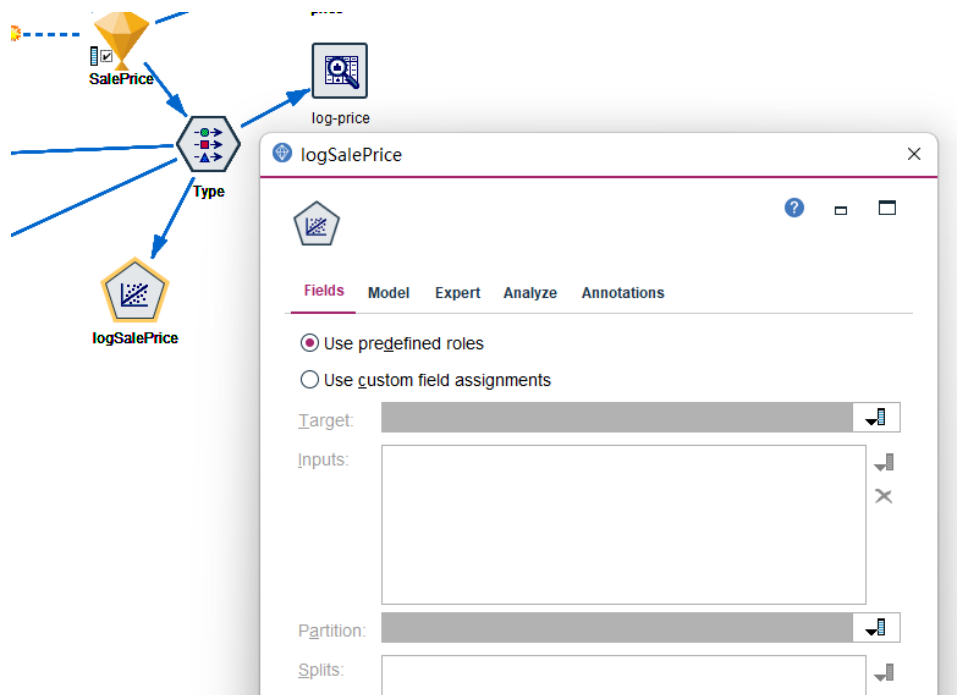


Figure6.3.2: Multivariate Regression model Parameter Settings

6.3.3 Random Forest model

For the random forest model, we need to specify the input and output variables. Similar to the multivariate regression model, our input variables are all explanatory variables (filtered earlier), and the output explanatory variable is the log-scale house price (AUD).

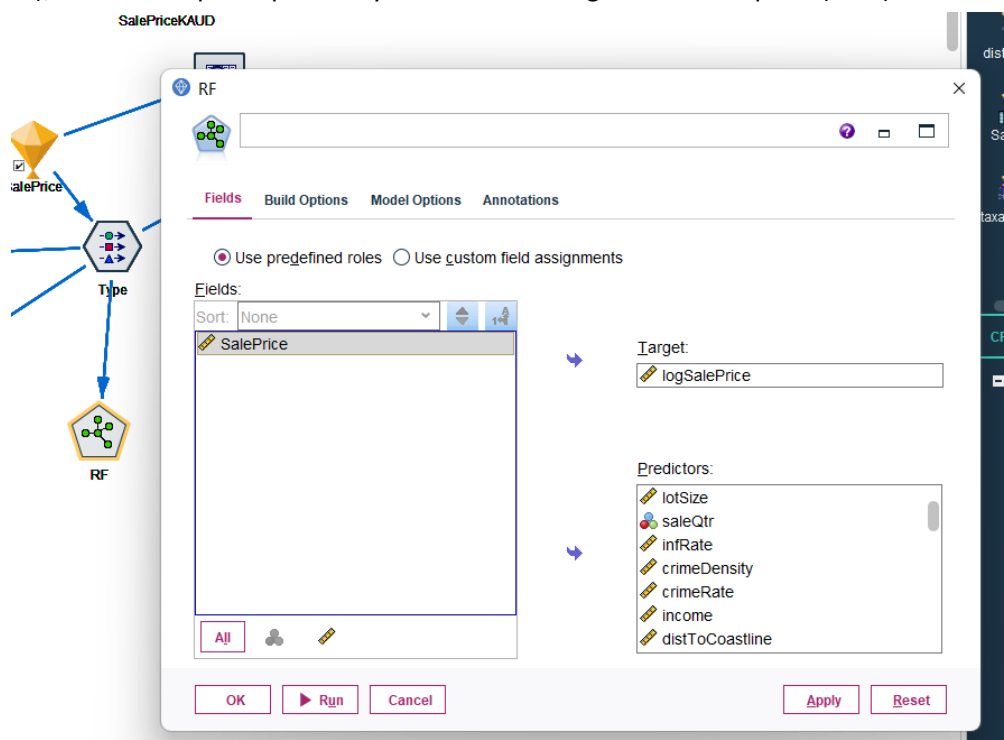


Figure6.3.3: Random Forest model Parameter Settings

We also need to set the number of trees used in the bagging process. I set it to 10 here, which is also the default setting in the SPSS modeler.

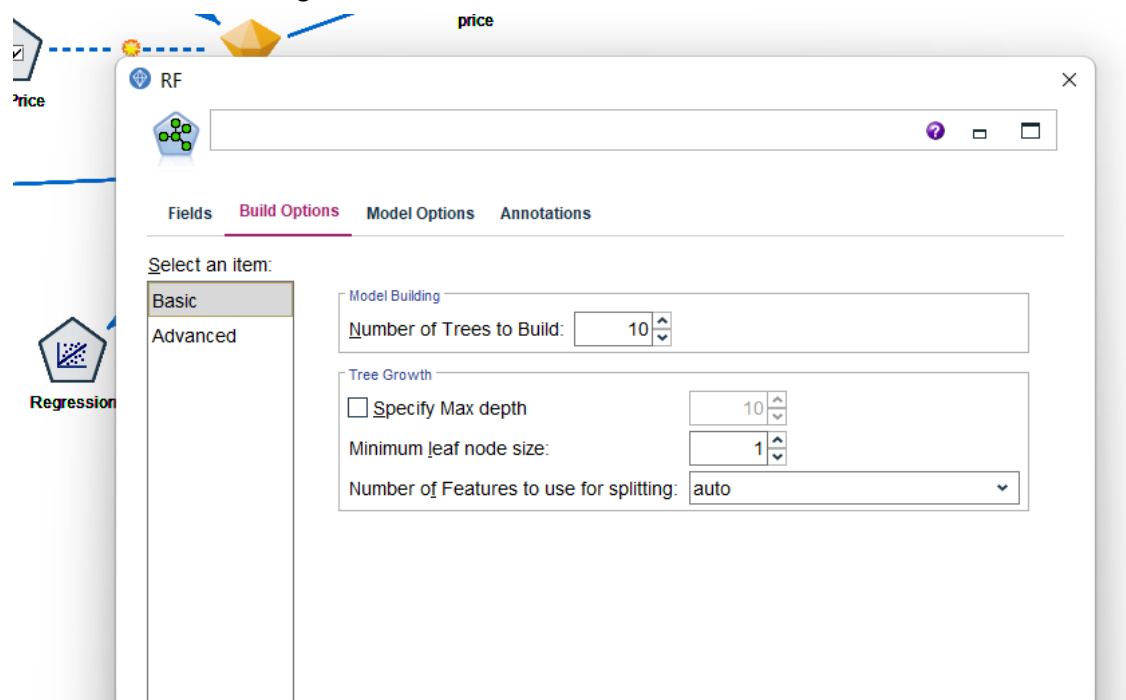


Figure6.3.4: Random Forest model Parameter Settings

In order to get reproducible results, for algorithms with randomness we need to set a seed. Here I set it to 648, which is the last three digits of my UPI.

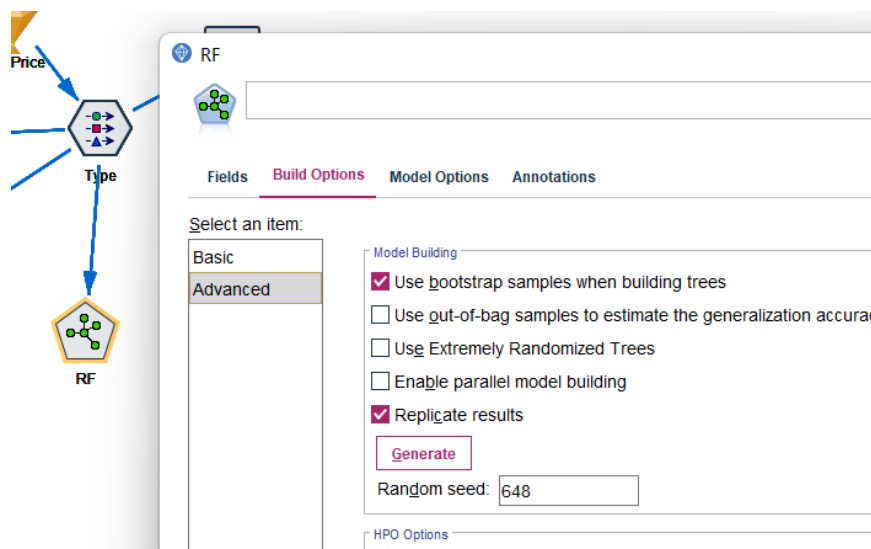


Figure6.3.5: Random Forest seeds

So far, we have completed the construction and parameter setting of the models. In the next section, we'll run these models.

7. Data Mining

7.1 Create and justify test designs

In statistical modeling practice, some methods of data partitioning are often used. This is based on the Golden rule of machine learning: – The test data cannot influence training the model in any way. [9] Therefore, we need to separate the training and test sets. Models are trained on the training set and evaluated on the test set. Usually we only care about the performance of the model on the test set, because over-optimizing on the training set can lead to overfitting, which can lead to large variance. In supervised learning, overfitting happens when our model captures the noise along with the underlying pattern in data. [10] To a large extent, the process of machine learning is a Bias-Variance Tradeoff process.

Specifically, we can use a 70/30 split, that is, use 70% of the data as the training set and the remaining 30% as the test set. This is a less computationally expensive solution that can be used when the amount of data is sufficient. This method is also known as Hold-out. Another popular solution is cross-validation, which is suitable for situations with a small amount of data, but has a high computational cost. Cross-validation randomly splits the dataset into "k" groups. One of the groups is used as the test set, and the remaining groups are used as the training set. The model is trained on the training set and scored on the test set. The process is then repeated until each unique group is used as a test set. [11]

Determining the size of the training/testing set is also a trade-off. A larger training set would allow our model to be trained on a larger dataset, but a correspondingly smaller test set would lead to less reliable test performance. Here we'd like to reserve a relatively large independent test set (perhaps be better called validation set as in some literature), so the prediction errors evaluated on the test set can be a bit more reliable. [12]

7.2 Conduct data mining – classify, regress, cluster, etc.

First, we need to implement the train/test split.

As previously described, the training set uses 70% of the data and the test set uses 30% of the data. Also, seed needs to be set up to produce reproducible results.

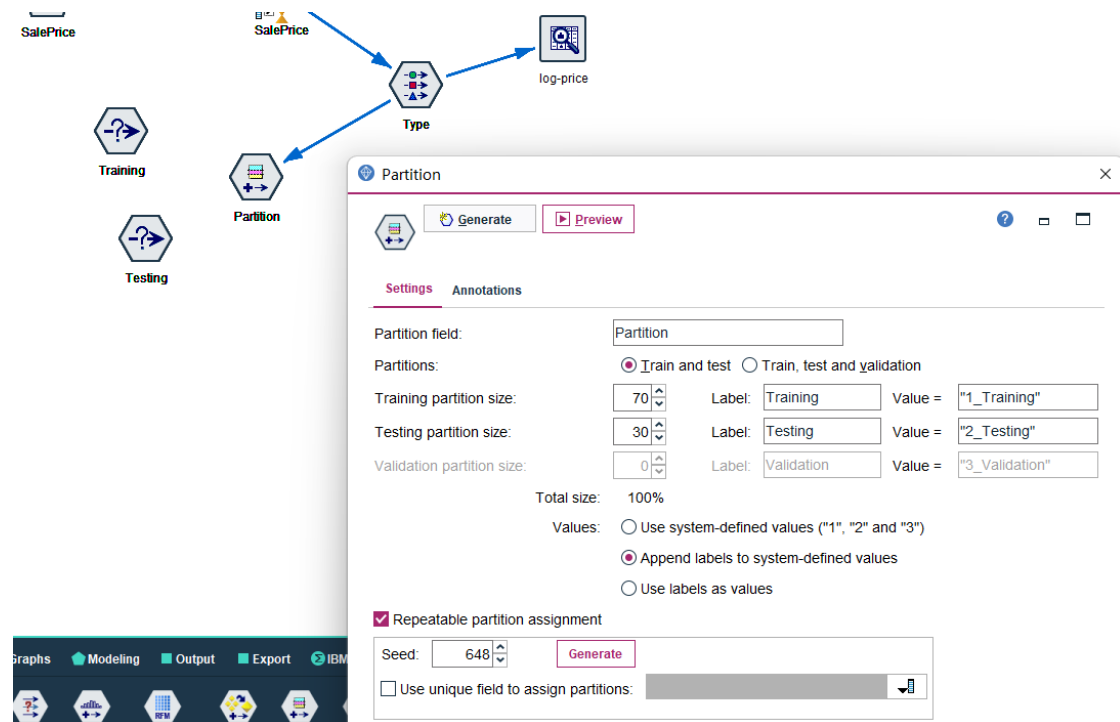


Figure 7.2.1: train/test split

There are 24593 rows in the training set and 10628 rows in the test set.

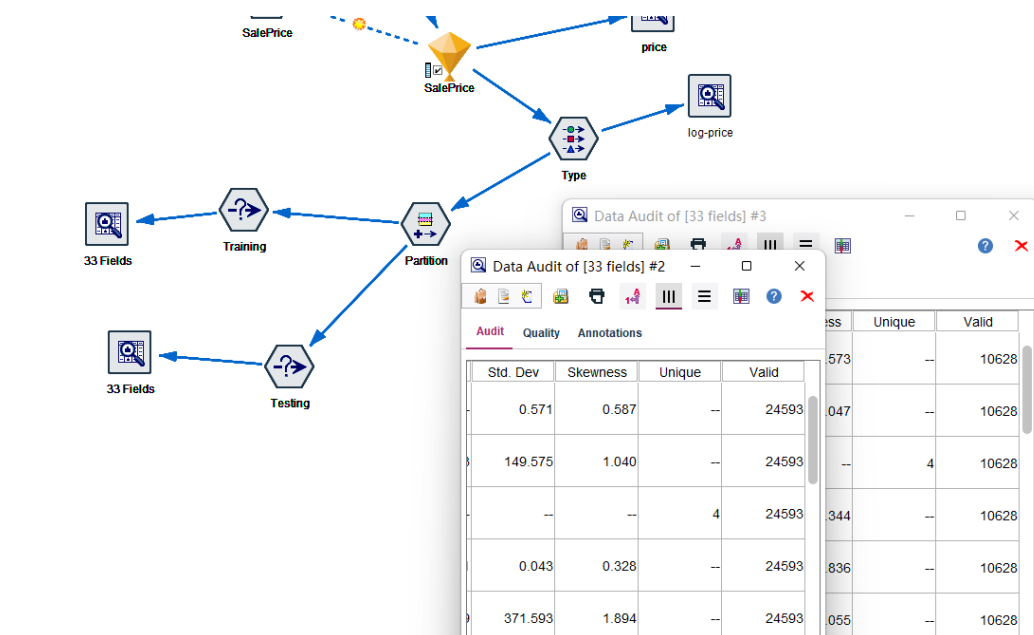


Figure 7.2.2: train/test split

I performed the data mining scheme described earlier in the SPSS Modeler: I fit three models, linear model, Multivariate Regression model and Random Forest on the training set, and verified the performance of the models on the test set. The core architecture of this part is shown in the figure:

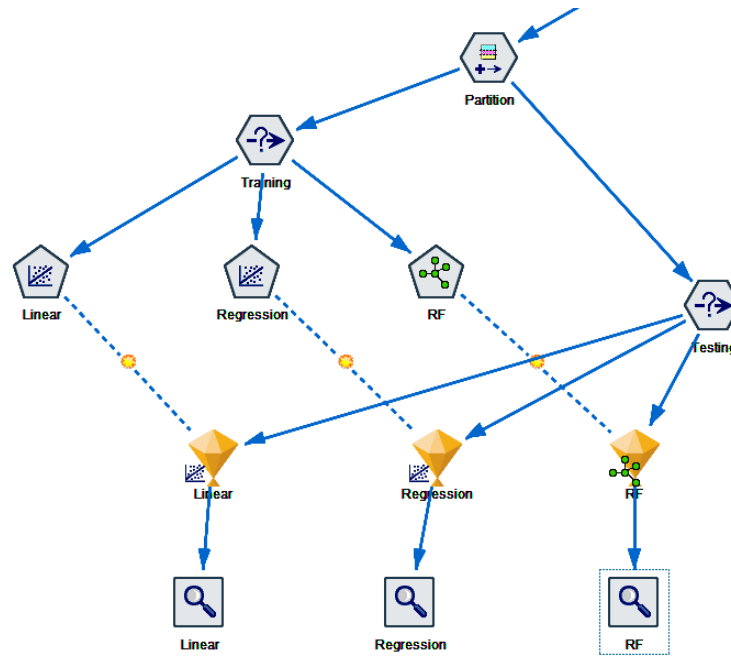


Figure7.2.3: core architecture

For regression models, the quality of the model can be measured using the metric "linear correlation". "Linear correlation" refers to the linear correlation between the predicted response variable and the actual response variable in the test set. Ideally, if the predicted value and the actual value are exactly the same, obviously the linear correlation will be 1. In other words, the larger the value of the linear correlation, the better the prediction performance of the model.

7.3 Search for patterns

From the results, the linear correlation of the linear model is the lowest at 0.343. This means that this model cannot predict accurately. The linear correlation of the multivariate regression model is 0.835, which is a great improvement over the linear model. The linear correlation of random forest is the largest at 0.861.

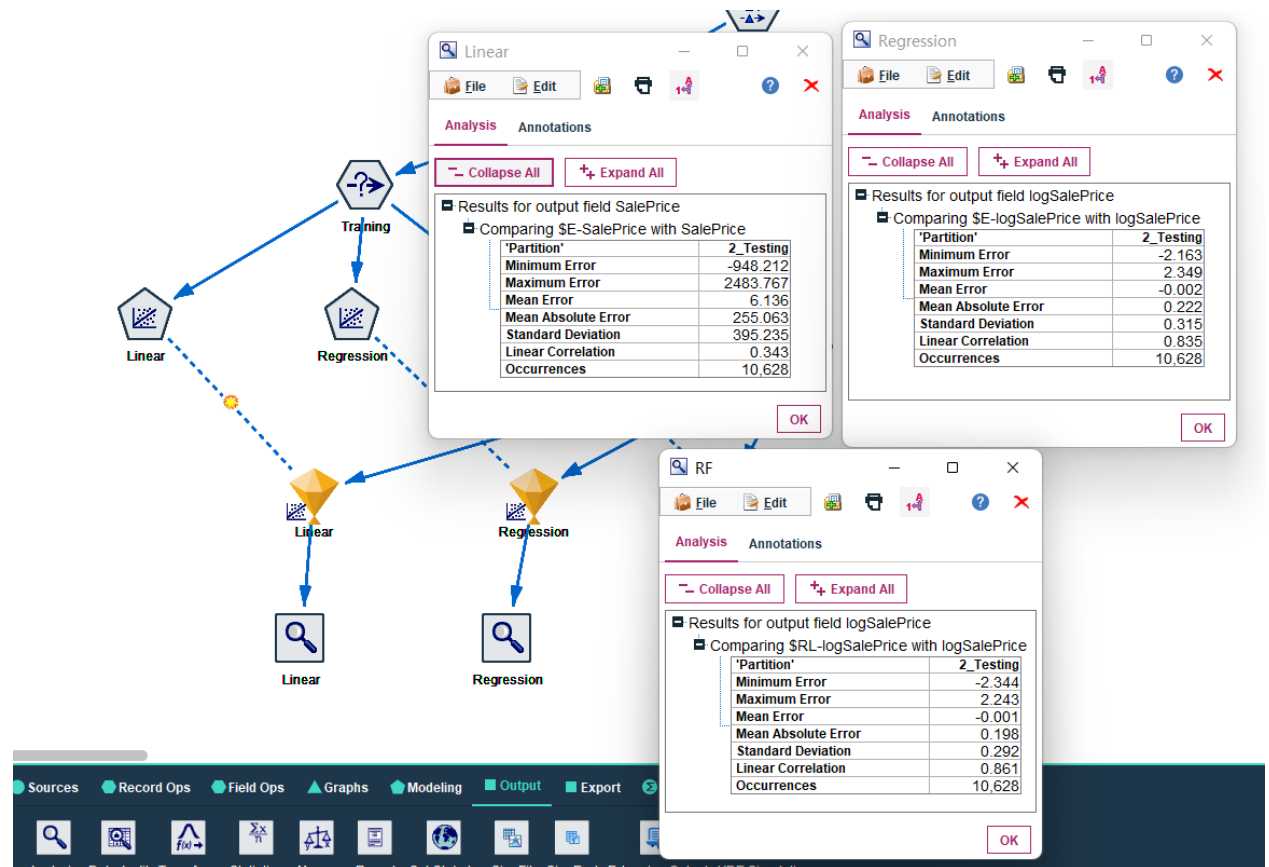


Figure7.3.1: Model performance

The prediction performance of the linear model is very low, while the prediction performance of the multivariate regression model has been greatly improved compared to linear model, and the prediction performance of the random forest is the strongest. This is not surprising: only one explanatory variable is used in the linear model, resulting in a model that is too simplistic and has a large bias in the model. This is a typical underfitting model, which will not perform well on either the training set or the test set. Both multivariate regression models and random forests use the full dataset as explanatory variables and thus have much better predictive performance. Among them, random forest uses a more complex algorithm, which has slightly higher prediction performance.

8. Interpretation

8.1 Study and discuss the mined patterns

In the previous steps, we identified a supervised method rather than an unsupervised method to achieve our data mining goals, and further confirmed that we used a regression method rather than a classification method.

As mentioned in Section 7.3, random forests have the best predictive performance. We will discuss it in detail in subsequent subsections.

Before fitting a model, it is necessary to determine whether the data conform to the statistical assumptions of the model. For example, a linear model assumes that the relationship between the response variable and the explanatory variable is linear, which is why the response variable I use when fitting a linear model is the property price rather than its logarithm.

On the other hand, when fitting multivariate regression models and random forests, you need to fit as many variables as possible to the model's assumptions. In the previous Data Audit node, we saw that many variables were linearly related to the log-scale property price, so the response variable I used when fitting the two models was the logarithm of the property price.

Finally, we used "linear correlation" to judge the predictive performance of the model. The closer the "linear correlation" is to 1, the closer the predicted value of the model is to the actual value in the test set, and the better it can describe the trend of the response variable, so the better the model is.

8.2 Visualize the data, results, models, and patterns

In the predicted value -- actual value plot of the linear model, it can be clearly seen that the linear relationship between the two is not very obvious. The scattered distribution of data points is shown in the plot.

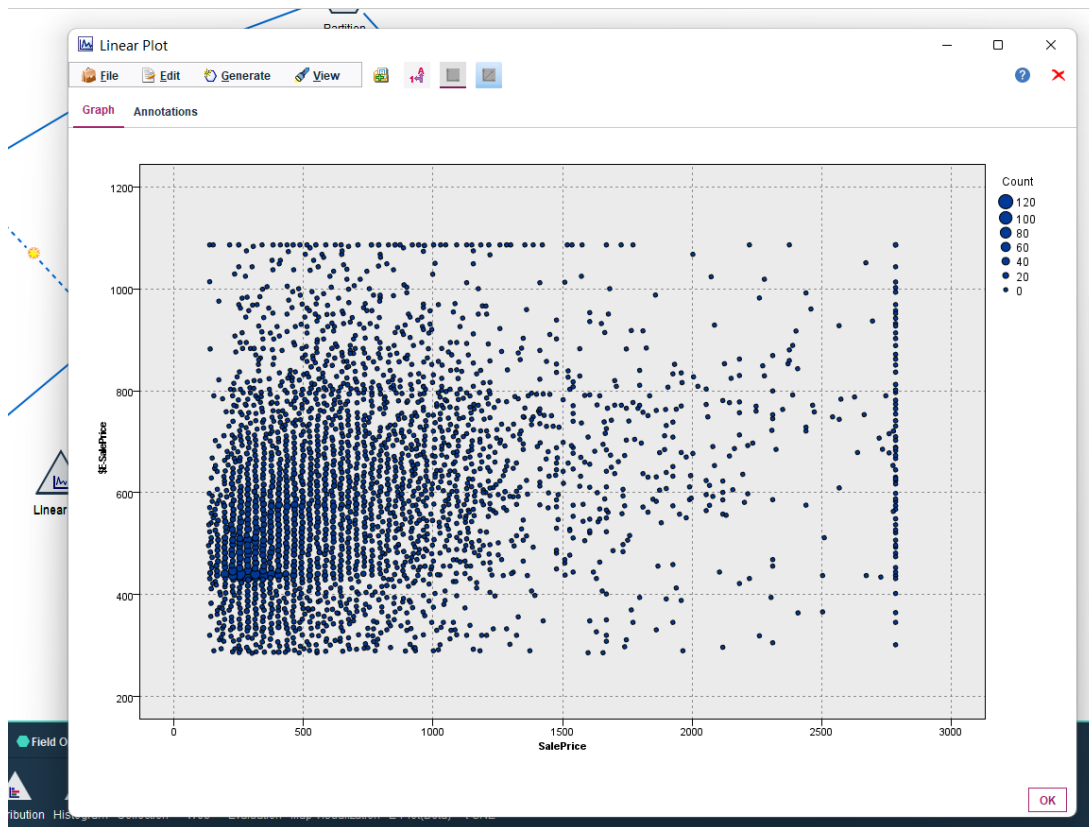


Figure8.2.1: performance of linear model

In the plot of the multivariate regression model, a clear linear relationship can be observed.

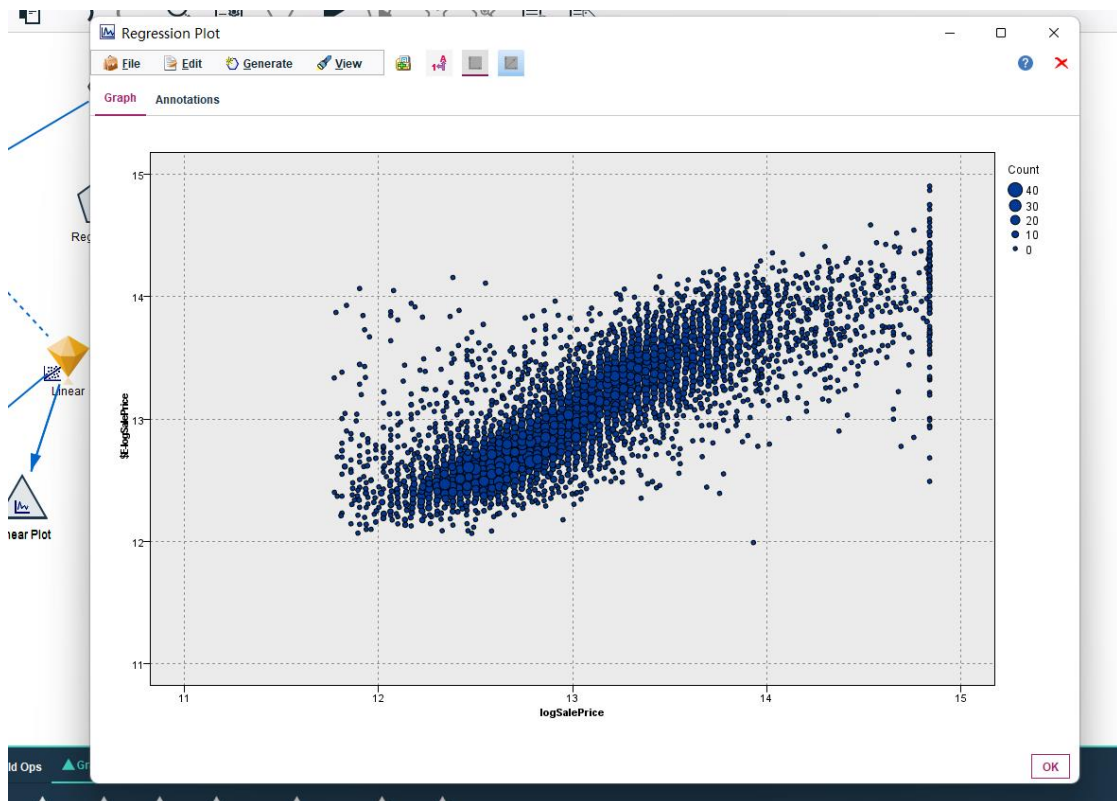


Figure8.2.2: performance of regression model

Random forest is similar to multivariate regression model.

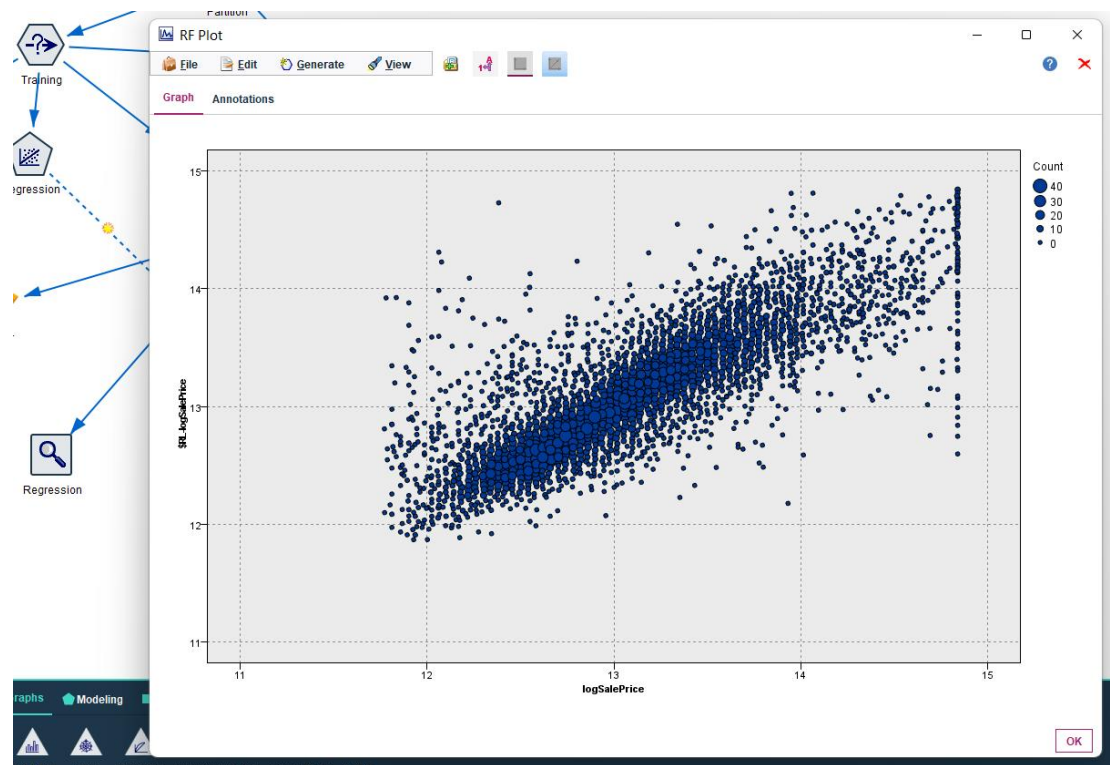
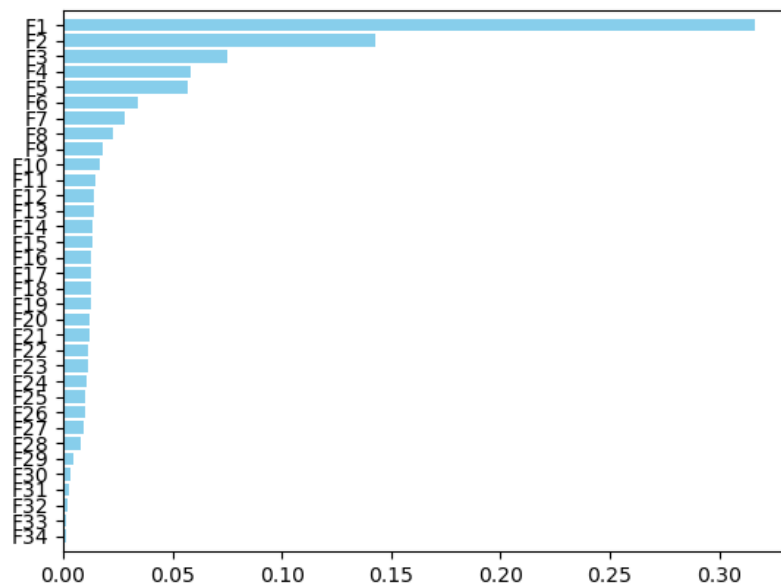


Figure8.2.3: performance of RF model

Random forest gives the importance ranking of explanatory variables.

For housing prices, the top four most important influencing factors are distance from house to the Sydney Harbour Tunnel, income, local SO2 concentration and lot size.

Random Forest Model



Features names for short

Original field name	Field name on graphic
distToTunnel	F1
income	F2
SO2	F3
lotSize	F4

Figure8.2.4: feature importance

Some variables are positively correlated with the logarithm of property prices, such as income level; others are negatively correlated with the logarithm of property prices, such as the distance from the property to the coastline.

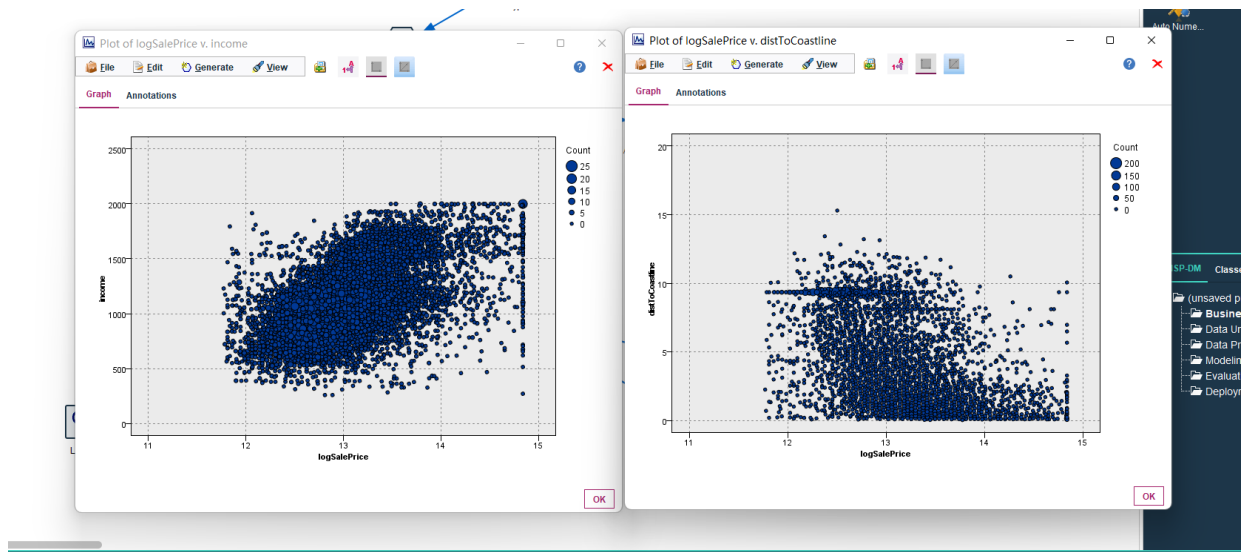


Figure8.2.5: Positive and Negative Correlation

8.3 Interpret the results, models, and patterns

The model parameters of the **Linear model** are as follows:

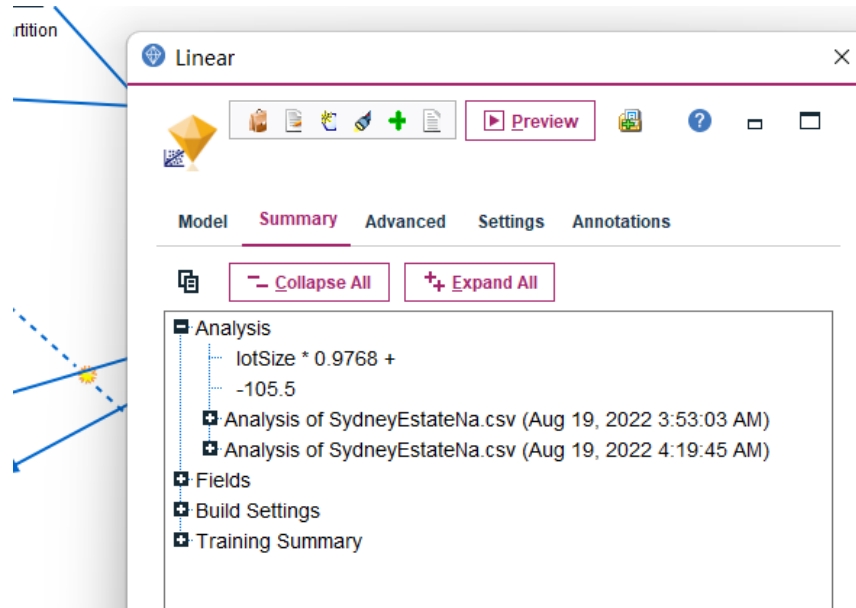


Figure8.3.1: Linear Model Parameters

From the model parameters of the linear model, it can be concluded that for every one square meter increase in the lot size of the property, the price of the property will increase by 0.9768 thousand AUD, or 976.8 AUD. This is a conclusion describing the unit price per square meter of a property.

The model parameters of the **Multivariate Regression model** are as follows:

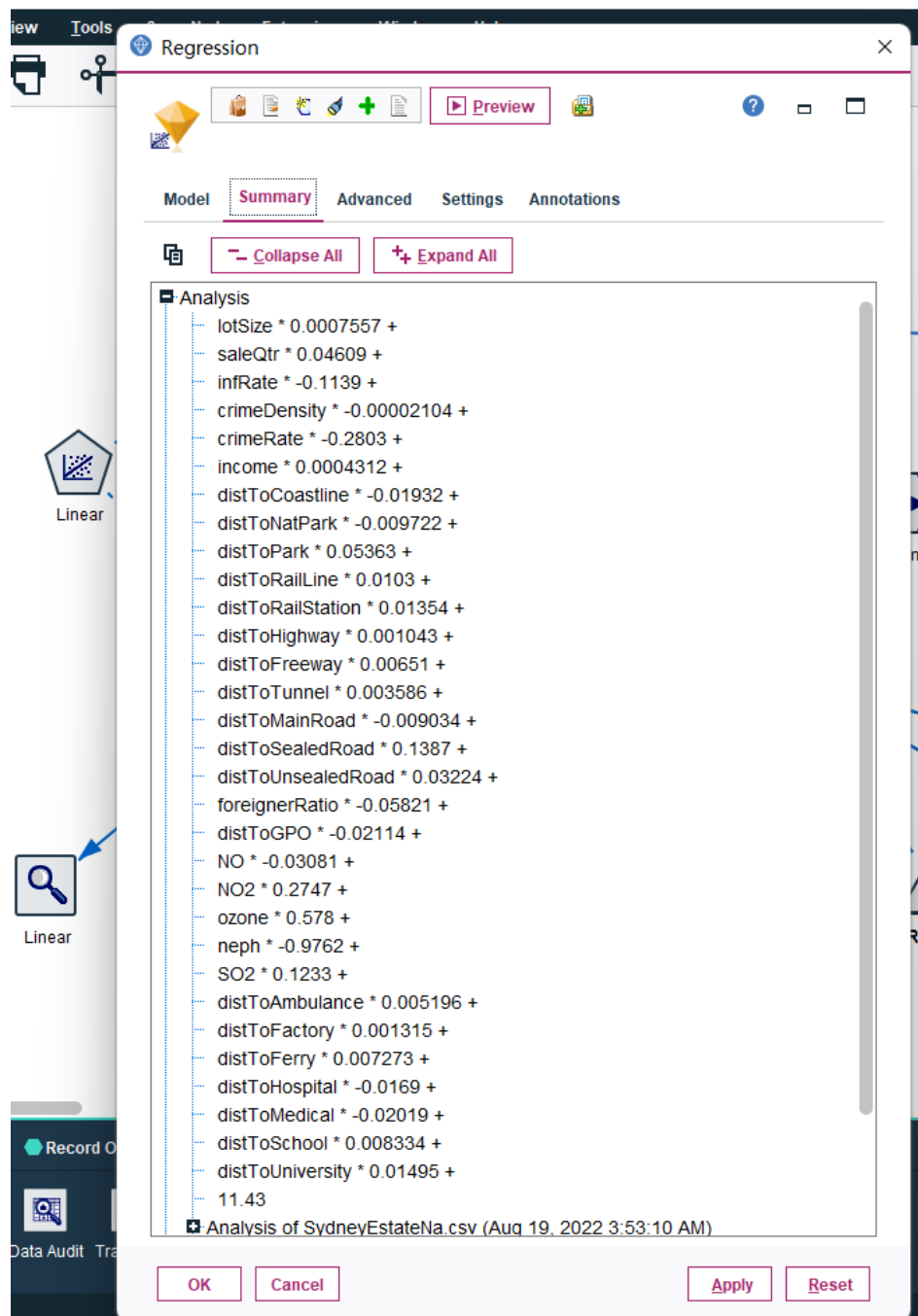


Figure8.3.2: Multivariate Regression Model Parameters

This model is a bit harder to interpret because the response variable is on the log scale.

For example, if you want to explain the relationship between lot size and house price, you can say that $\log(\text{Price})$ increases by 0.0007557 for every 1 square meter increase in lot size, that is, property price multiplied by $\exp(0.0007557) = 1.000756$. This is obviously problematic because the relationship between property price and lot size is not exponential. This model is more predictive, but less interpretable.

However, we can still see some trends from the parameter list. When the coefficient of an

explanatory variable is greater than 0, it can be said to be positively correlated with the response variable, such as lot size and income; when the coefficient of an explanatory variable is less than 0, it can be said to be negatively correlated with the response variable, such as crime rate and the distance to the coastline. This is not surprising, intuitively, houses with larger footprints are more expensive; people with higher incomes tend to buy more expensive houses; neighborhoods with high crime rates lead to lower prices; people are more Prefer to live closer to the coastline, so the farther from the coastline, the cheaper the property.

8.4 Assess and evaluate results, models, and patterns

We fit a total of three models: linear model, multiple regression model and random forest. The performance of the linear model is the worst, with severe underfitting. In actual forecasting, we cannot use such a model. Multiple regression and random forests perform similarly, with random forests slightly better than multiple regression models. However, the running time of the random forest model is significantly longer, and the cost of slightly better prediction performance is obvious. In actual data mining practice, data mining goals should be fully considered, and whether more complex, longer-running, and harder-to-interpret models should be used for small improvements. My answer tends to be no.

From the parameter coefficient of the multivariate linear model, the factors that lead to higher real estate prices are:

--larger lot size

-- Higher average weekly income of the suburb in which the house is located

--Greater distance from house to the nearest railway line

.....

Factors that lead to lower property prices include:

-- farther from the coastline

-- Higher proportion of foreigners in the suburb

--Higher crime rate measure for the suburb

.....

The multiple linear model also gave us some incredible conclusions, such as the higher the concentration of nitrogen dioxide level around the property, the higher the property price. Nitrogen dioxide is a pollutant and people should avoid living in areas with high nitrogen dioxide levels.

8.5 Iterate prior steps (1 – 7) as required

In Section 7.3, I compared the performance of the three models I built earlier. SPSS Modeler provides a way to combine different models and end up with an ensemble model that generally has better performance on the test set than any individual model.

I generated such an ensemble model on the training set using the Auto Numeric node.

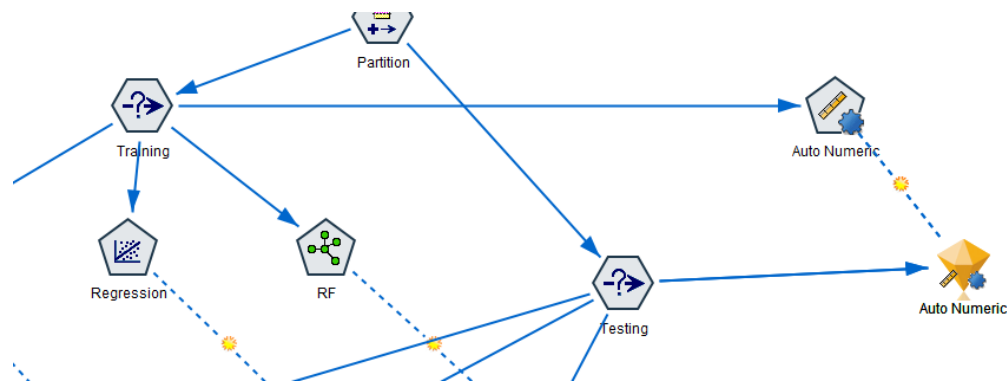


Figure8.5.1: Auto Numeric and ensemble model

Check the predictive performance of the ensemble model using the analysis node:

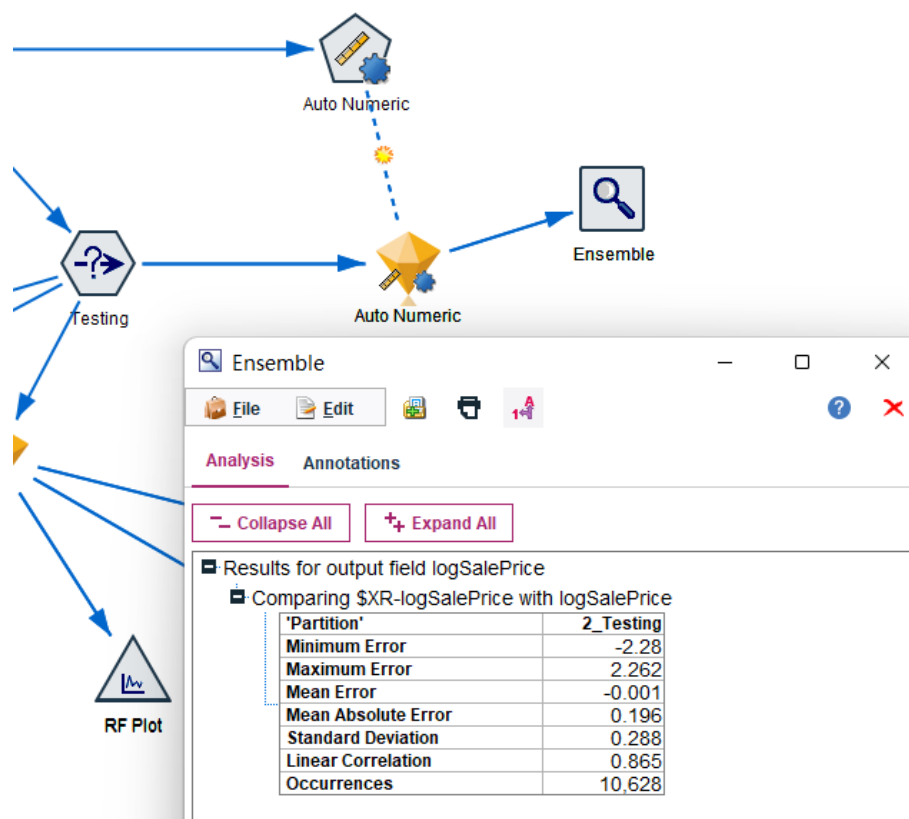


Figure8.5.2: Model performance

The linear correlation of this ensemble model is 0.865, while the previous best model (random forest) is 0.861. Therefore, this ensemble model does have higher predictive performance. I think the idea of this ensemble model is somewhat similar to bagging (e.g. random forest).

References

- [Aut], J. D. H. R. (2021a, November 23). *HRW: Datasets, Functions and Scripts for Semiparametric Regression Supporting Harezlak, Ruppert & Wand (2018)*. R Package Documentation. <https://rdr.io/cran/HRW/>
- [Aut], J. D. H. R. (2021b, November 23). *Sydney real estate*. R Package Documentation. <https://rdr.io/cran/HRW/man/SydneyRealEstate.html>
- Realestate, E. (2021, October 20). *Factors That Affect Housing Prices in Sydney*. Etch Real Estate. <https://www.etchrealestate.com.au/factors-that-affect-housing-prices-in-sydney/>
- Sustainable cities and human settlements* | Department of Economic and Social Affairs. (2018). United Nations. <https://sdgs.un.org/topics/sustainable-cities-and-human-settlements>
- Yee, T. (Ed.). (2022). Introduction to Data Mining. In *STATS 784: Statistical Data Mining* (pp. 60–67).
- Cialdella, L. (2020, August 30). *When do we log transform the response variable? Model assumptions, multiplicative combinations and log-linear models*. Casual Inference. <https://lmc2179.github.io/posts/multiplicative.html>
- Agrawal, P., Gupta, C., Sharma, A., Madaan, V., & Joshi, N. (2022). *Machine Learning and Data Science: Fundamentals and Applications* (1st ed.). Wiley-Scrivener.
- Education, I. C. (2021, January 26). *Random Forest*. IBM. <https://www.ibm.com/cloud/learn/random-forest>
- Machine Learning and Data Mining. (2016). In *CPSC 340 UBC* (p. 19).
- Singh, S. (2022, February 18). *Understanding the Bias-Variance Tradeoff - Towards Data Science*. Medium. <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>
- Allibhai, E. (2022, June 21). *Hold-out vs. Cross-validation in Machine Learning - Eijaz Allibhai*. Medium. <https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f>
- Prasad, A. (2022, January 6). *Regression Trees | Decision Tree for Regression | Machine Learning*. Medium. Retrieved September 23, 2022, from <https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047>
- Random forests - classification description. (n.d.). Retrieved September 23, 2022, from https://www.stat.berkeley.edu/%7Ebreiman/RandomForests/cc_home.htm
- Supervised vs. Unsupervised Learning: What's the Difference? (2021, March 12). IBM. Retrieved September 23, 2022, from <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>
- Gradient boosting. (2022, September 1). Wikipedia. Retrieved September 23, 2022, from https://en.wikipedia.org/wiki/Gradient_boosting

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright.

(See: <https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html>).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."