2020 MSA Phase 1 – Data Pathway – Jordan Kim

Executive Summary

The given dataset is addresses in Auckland. The dataset includes details about the address, such as, the amount of rooms, location, age group of people in the area.

The final analysis will be based on 1049 observations on 12 different variables. The response variable will be CV, which stands for the capital value of property. This value is used to calculate payable rates and approximates house value.

After exploring the dataset, through summaries about the data, and creating visuals, we can obtain a better understanding. We have also created a linear model for this dataset.

Initial Data Exploration

	Bedrooms	Bathrooms	Address	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Suburbs
0	5	3.0	106 Lawrence Crescent Hill Park, Auckland	714	960000	-37.012920	174.904069	7009770	48	27	24	21	24	21	Manurewa
1	5	3.0	8 Corsica Way Karaka, Auckland	564	1250000	-37.063672	174.922912	7009991	42	18	12	21	15	30	Karaka
2	6	4.0	243 Harbourside Drive Karaka, Auckland	626	1250000	-37.063580	174.924044	7009991	42	18	12	21	15	30	Karaka
3	2	1.0	2/30 Hardington Street Onehunga, Auckland	65	740000	-36.912996	174.787425	7007871	42	6	21	21	12	15	Onehunga
4	3	1.0	59 Israel Avenue Clover Park, Auckland	601	630000	-36.979037	174.892612	7008902	93	27	33	30	21	33	Clover Park

This was the given dataset. Then we added two new columns as the assignment states. One for population count and another for Deprivation Index.

	Bedrooms	Bathrooms	Address	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Suburbs	C18_CURPOP	NZDep2018
0	5	3.0	106 Lawrence Crescent Hill Park, Auckland	714	960000	-37.012920	174.904069	7009770	48	27	24	21	24	21	Manurewa	174	6.0
1	5	3.0	8 Corsica Way Karaka, Auckland	564	1250000	-37.063672	174.922912	7009991	42	18	12	21	15	30	Karaka	129	1.0
2	6	4.0	243 Harbourside Drive Karaka, Auckland	626	1250000	-37.063580	174.924044	7009991	42	18	12	21	15	30	Karaka	129	1.0
3	2	1.0	2/30 Hardington Street Onehunga, Auckland	65	740000	-36.912996	174.787425	7007871	42	6	21	21	12	15	Onehunga	120	2.0
4	3	1.0	59 Israel Avenue Clover Park, Auckland	601	630000	-36.979037	174.892612	7008902	93	27	33	30	21	33	Clover Park	231	9.0

If we have a closer look at the land area, towards the bottom of the dataset, the values are not consistent. There are units written in some of these rows, like shown below.

1046	4	1.0	19 Landscape Road, Auckland	1368 m²	670000	-36.899255	174.761165	7005464	54	18	15	24	21	27	Mount Eden	159	1.0
1047	6	1.0	56 Galway Street, Auckland	607 m²	1200000	-36.844933	174.770001	7005497	15	27	24	15	18	30	Auckland Central	129	6.0
1048	5	3.0	28A Hayr Road, Auckland	453 m²	1250000	-36.912242	174.756726	7007758	36	30	45	21	24	21	Three Kings	180	6.0
1049	5	2.0	27 Market Road, Auckland	1854 m²	5300000	-36.879665	174.787668	7005745	48	18	12	15	36	45	Remuera	174	1.0
1050	3	1.0	23 William Avenue, Auckland	806 m²	1665000	-36.897104	174.800171	7005917	54	33	27	27	15	30	Greenlane	192	4.0

We can remove this with the use of the following code whilst converting this attribute to a float:

```
dfFinal["Land area"] = dfFinal["Land area"].str.extract('(\d+)').astype(float)
```

We then do another check on the variable types and see that Bathrooms is a float, where you'd expect it to be an integer.

```
dfFinal.dtypes
```

```
Bedrooms         int64
Bathrooms      float64
Land area      float64
CV               int64
Latitude       float64
Longitude      float64
SA1              int64
0-19 years       int64
20-29 years      int64
30-39 years      int64
40-49 years      int64
50-59 years      int64
60+ years        int64
C18_CURPOP       int64
NZDep2018      float64
dtype: object
```

We find the unique values for this attribute and find there are rows filled with NaN. We can observe that only 2 rows have NaN values. I decided to remove these two rows as it is a very small portion from the current dataset. After removing those two rows we end up with a data frame of 1049 observations.

```
dfFinal["Bathrooms"].unique()
```

```
array([ 3.,  4.,  1.,  2.,  5., nan,  6.,  8.,  7.])
```

```
dfFinal["Bathrooms"].isna().sum()
#dfFinal.isnull().values.any()
```

```
2
```

```
dfFinal = dfFinal[dfFinal["Bathrooms"].notna()]
```

```
dfFinal.shape
```

```
(1049, 15)
```

We also drop the non-numerical attributes being address and suburb and end up with our final data frame's description, shown below.

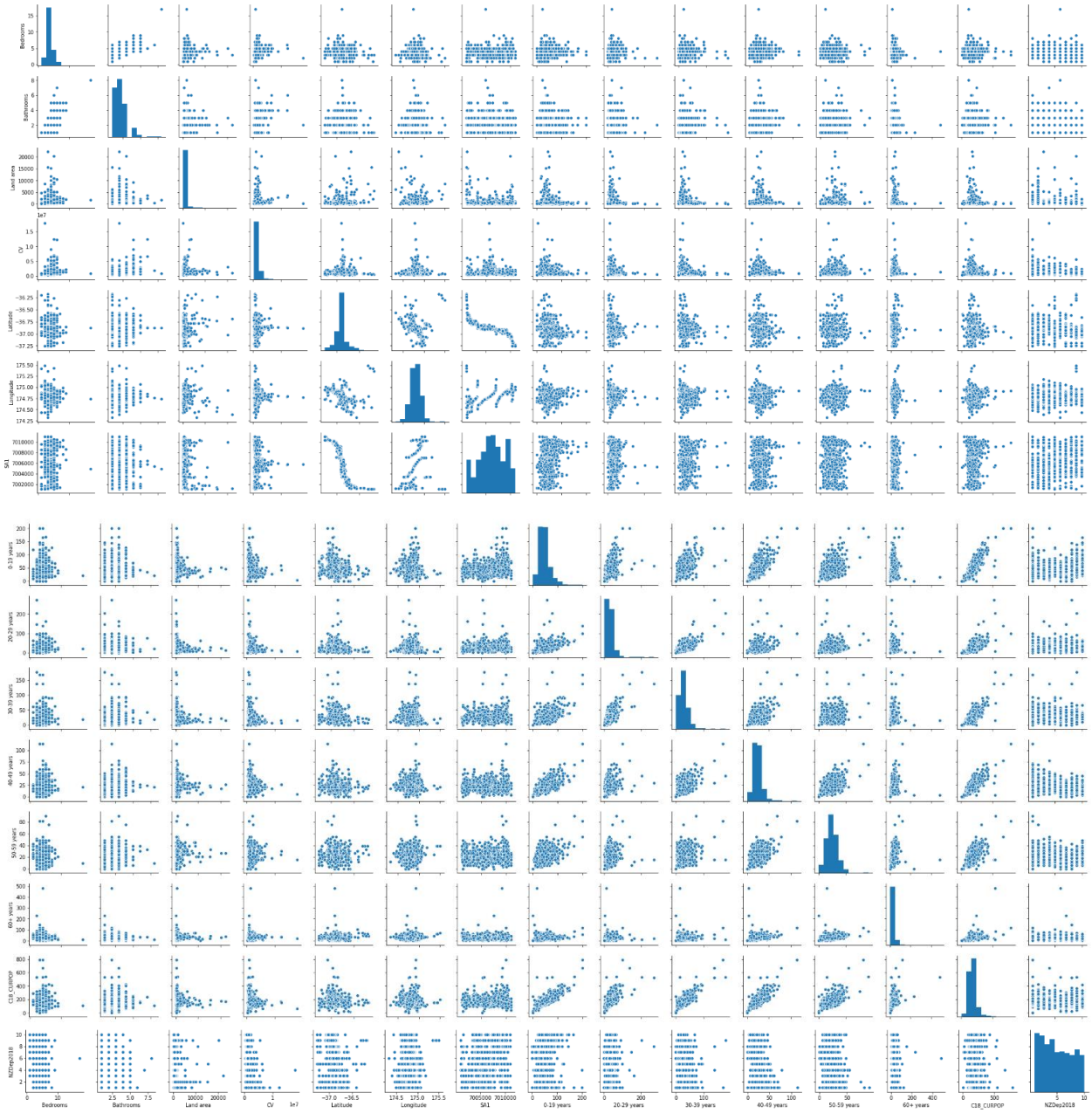
```
dfFinal.describe()
```

	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	C18_CURPOP	NZDep2018
count	1049.000000	1049.000000	1049.000000	1.049000e+03	1049.000000	1049.000000	1.049000e+03	1049.000000	1049.000000	1049.000000	1049.000000	1049.000000	1049.000000	1049.000000	1049.000000
mean	3.776930	2.073403	858.185891	1.387926e+06	-36.893897	174.799615	7.006327e+06	47.525262	28.893232	26.979981	24.125834	22.612965	29.382269	179.776930	5.069590
std	1.170487	0.992985	1589.433957	1.184027e+06	0.130158	0.119468	2.587674e+03	24.709758	20.995139	17.934747	10.953205	10.220137	21.820173	71.057174	2.913171
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000	1.000000
25%	3.000000	1.000000	323.000000	7.800000e+05	-36.950722	174.722474	7.004424e+06	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000	2.000000
50%	4.000000	2.000000	572.000000	1.080000e+06	-36.893368	174.798648	7.006333e+06	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	174.000000	5.000000
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.856192	174.880945	7.008385e+06	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	207.000000	8.000000
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	789.000000	10.000000

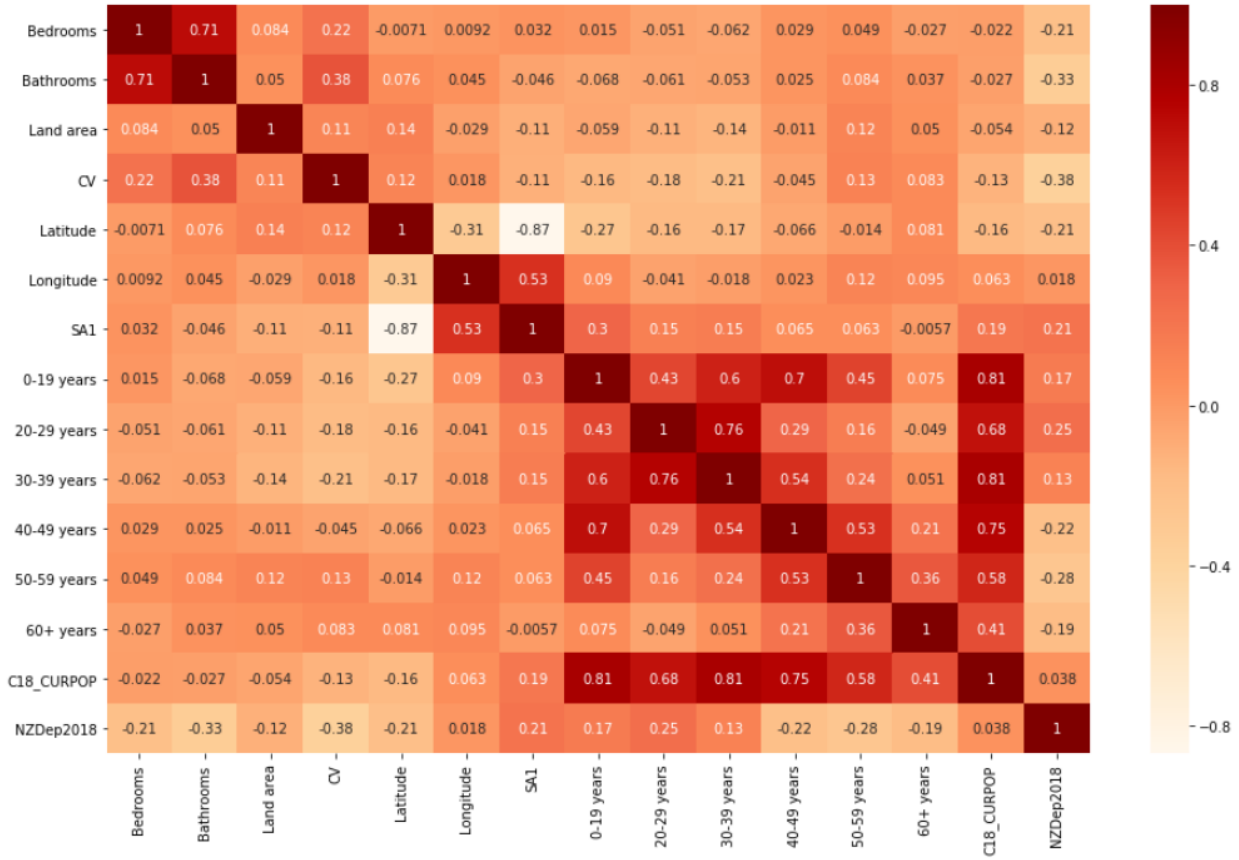
Correlation and Relationship

Numeric Relationships

The correlation between the numeric columns were calculated and observed in the below plots. Firstly, we have a pair plot

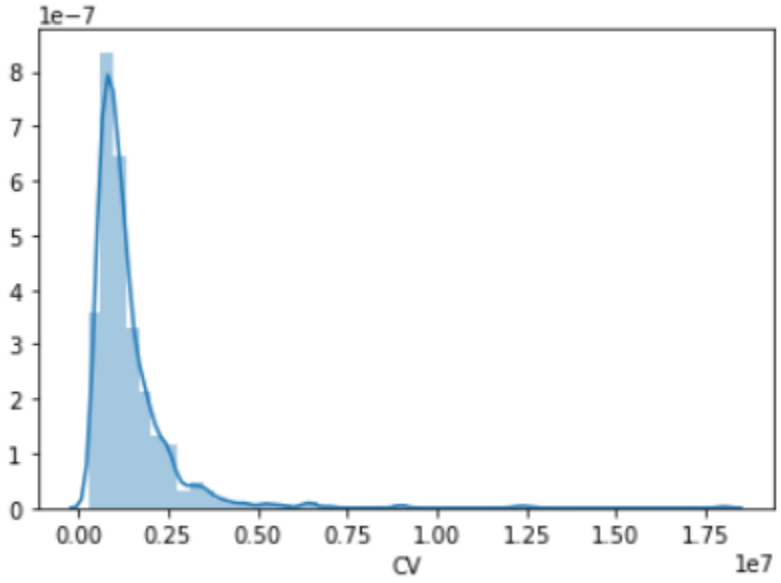


Secondly, we have a confusion matrix.



Analysis

We initially have a look at our response variable CV and create a simple histogram.



From this histogram, we can see that it is right skewed therefore it will be logged. After this, we drop 3 variables SA1, Latitude and Longitude. We decided to drop SA1 as it is just a classification and latitude and longitude as there isn't enough data to pick up the locations that generally have higher house prices, such as places with ocean view. Then we created a linear regression model with a 0.3 test, train split.

We end up with a model score of 0.3831. From this given value we can see that our data doesn't fit our model well. This could because our decision in picking linear regression is not suitable for our data or even including variables that are unrelated.

```
model.score(test_x, test_y)
```

```
0.3831047527965265
```

Conclusion

This analysis has shown that from our given data and after adding new columns, it cannot confidently predict the CV. This is shown through the low accuracy of our model which is 38%.