Fundamentals of Data Science

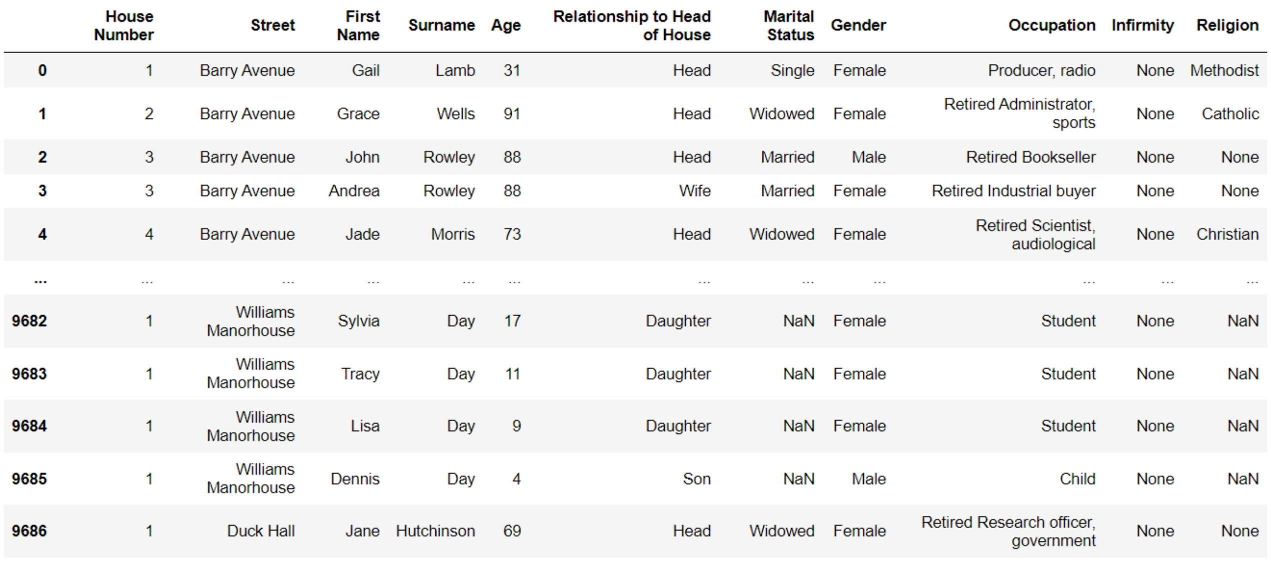
Project Report

Objective

This report sets out the household density and people's information in the UK and establishes the relationship between data tables and several graphs.

Moreover, this experiment analysis aims to examine what things should be developed in the land in the future and includes several ways how decision making to evaluate investment options can be achieved.

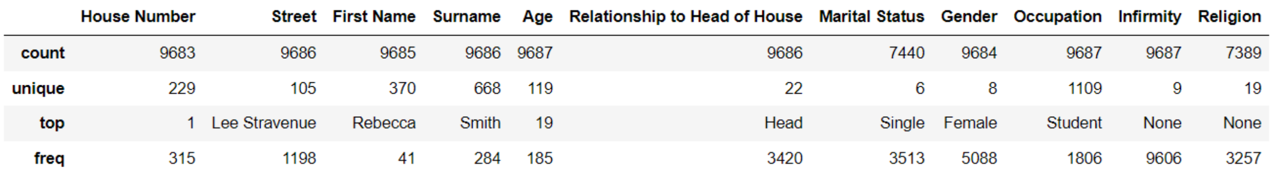
Data exploration Table 1



	House Number	Street	First Name	Surname	Age	Relationship to Head of House	Marital Status	Gender	Occupation	Infirmity	Religion
0	1	Barry Avenue	Gail	Lamb	31	Head	Single	Female	Producer, radio	None	Methodist
1	2	Barry Avenue	Grace	Wells	91	Head	Widowed	Female	Retired Administrator, sports	None	Catholic
2	3	Barry Avenue	John	Rowley	88	Head	Married	Male	Retired Bookseller	None	None
3	3	Barry Avenue	Andrea	Rowley	88	Wife	Married	Female	Retired Industrial buyer	None	None
4	4	Barry Avenue	Jade	Morris	73	Head	Widowed	Female	Retired Scientist, audiological	None	Christian
...	...	...	...	...	...	...	...	...	...	...	...
9682	1	Williams Manorhouse	Sylvia	Day	17	Daughter	NaN	Female	Student	None	NaN
9683	1	Williams Manorhouse	Tracy	Day	11	Daughter	NaN	Female	Student	None	NaN
9684	1	Williams Manorhouse	Lisa	Day	9	Daughter	NaN	Female	Student	None	NaN
9685	1	Williams Manorhouse	Dennis	Day	4	Son	NaN	Male	Child	None	NaN
9686	1	Duck Hall	Jane	Hutchinson	69	Head	Widowed	Female	Retired Research officer, government	None	None

To begin with, this data shows there are 11 categories in columns: House Number, Street, First Name, Surname, Age, Relationship to Head of House, Marital Status, Gender, Occupation, Infirmity, and Religion. Moreover, there are 9687 rows containing people’s information.

Problems with data Table 2



	House Number	Street	First Name	Surname	Age	Relationship to Head of House	Marital Status	Gender	Occupation	Infirmity	Religion
count	9683	9686	9685	9686	9687	9686	7440	9684	9687	9687	7389
unique	229	105	370	668	119	22	6	8	1109	9	19
top	1	Lee Stravenue	Rebecca	Smith	19	Head	Single	Female	Student	None	None
freq	315	1198	41	284	185	3420	3513	5088	1806	9606	3257

In a count row, shows many data in each category have missing values and incorrect data. Moreover, in a unique row, some data cannot plot graphs because there are many variables. Data cleaning should be used to make it better for analysis.

Data Cleaning

The reason why data cleaning is important is that some values are not correct for analysis and cannot be used to predict the future. There are a number of ways in which data can be cleaned. Five techniques are used in this report, as set out below.

1. Data should be standardized

In the category, house number, in Table 1, there is a problem. For example, there are nine in alphabet and 9 in numbers. However, the amount of numbers is more than the alphabet. As a result, nine should be changed to 9.

Moreover, the type of data should be correct. For example, the type of number should be an integer and the type of alphabet should be the string.

As a result, the first technique has been used in columns: House Number and Age.

2. Correcting false data

False data may be entered. For example, In the category religion, an entry of Jedi and Sith from Star Wars movies may occur (Lanham, 2016). It should be changed to irreligion.

On the other hand, in the age category, an entry of more than 122 years old may occur. The longest recorded human life is 122 years old (Collins ,2020). It can be changed by using step 3 (clean by comparing other columns).

As a result, the second technique has been used in columns: First Name, Surname, Age, Relationship to Head of House, Marital Status, Occupation, Infirmity and Religion.

3. Clean by comparing with other columns

Missing value: None, nan, and blank can be solved by using the relationship between house number and surname to find someone who forget to fill the street data. For example, in House Number 16 and Surname Mills. There are 4 people in this family. However, one person forgets to write the street. It can be filled by using the same street.

Similarly, it can be replaced by using mean mode or median. For instance, there is no value in an age of people’s information. It can solve the problem by using the median. In this data median is 35. Especially if, they are an employee it can be replaced by the median.

As a result, the third technique has been used in columns: House Number, Street, Surname, Age and Relationship to Head of House.

4. Resizing and grouping

The reason why data should be resized and grouped is that some categories have many variables. For instance, in occupation, there are 1109 variables. The specific descriptions, such as retired bookseller and air cabin crew, can be changed to retired and employee respectively.

As a result, the fourth technique has been used in columns: Marital Status, Gender, Relationship to Head of House and Occupation.

5. Law

Data may be entered that does not accord to UK law. For example, the head of the household must be more than 18 years old (eFile, 2021). It should be changed from whatever to “lodger”.

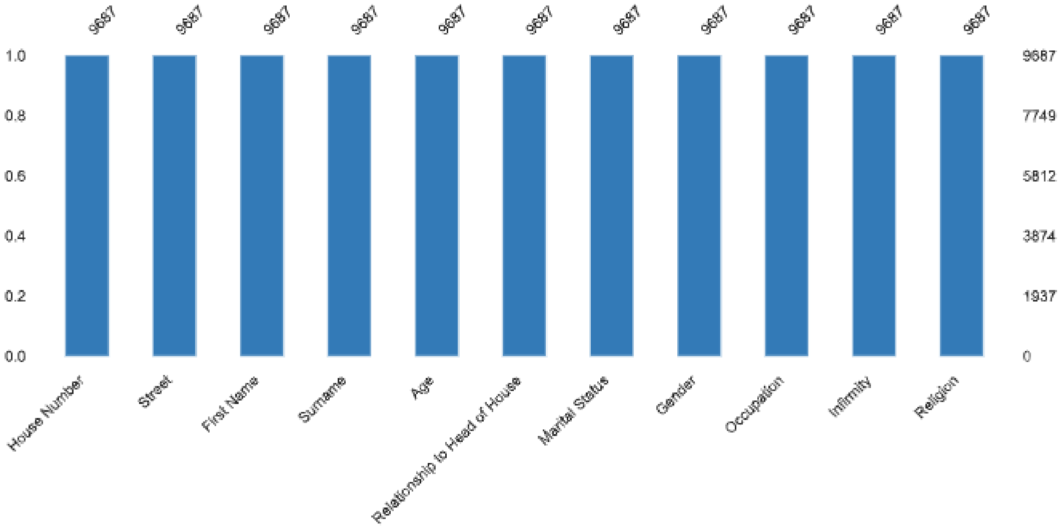
Similarly, people who are less than 16 years old cannot get married (Bloom, 2021). It should be changed from whatever to “single”.

As a result, the fifth technique has been used in columns: Relationship to Head of House and Marital Status.

Once these techniques have been applied, the data is ready for analysis.

Result of Data Cleaning

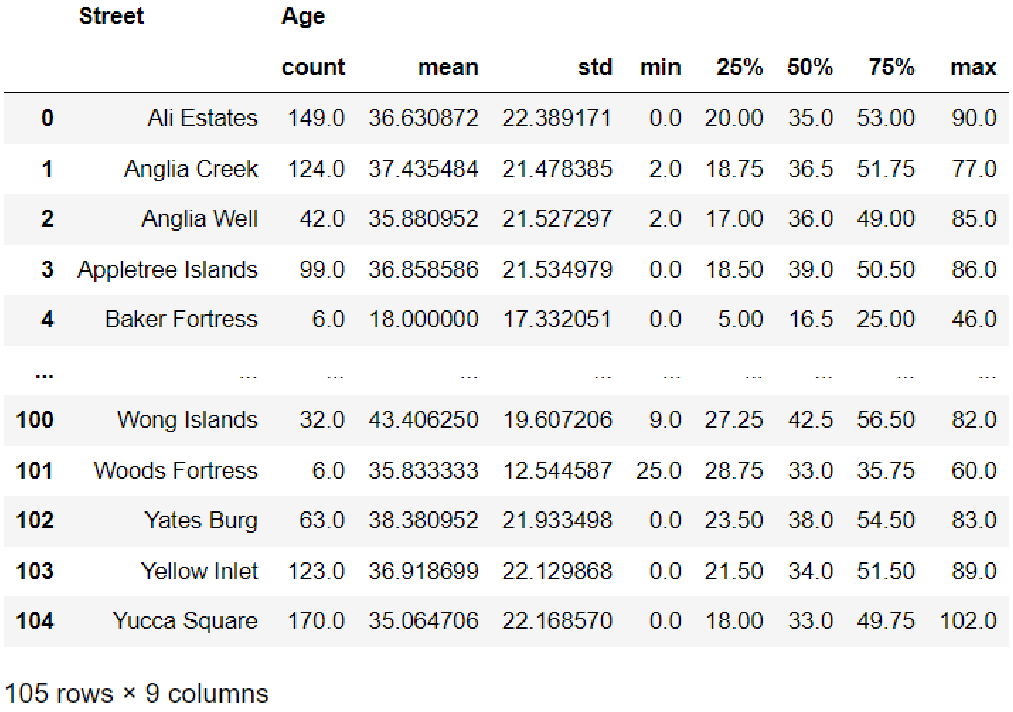
Figure number 1



Compared with the previous data table, this is more readable because the data has already been cleaned. All of the people's information in rows is 9687.

Household density

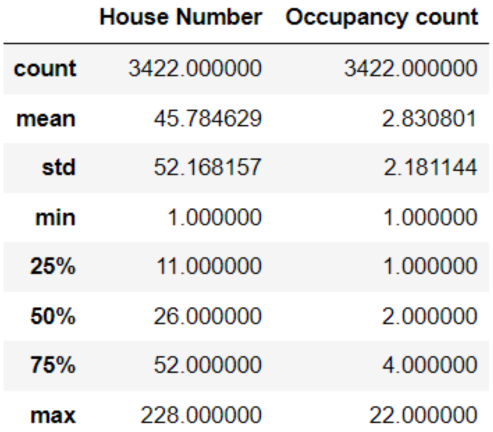
Table 3



	Street	Age							
		count	mean	std	min	25%	50%	75%	max
0	Ali Estates	149.0	36.630872	22.389171	0.0	20.00	35.0	53.00	90.0
1	Anglia Creek	124.0	37.435484	21.478385	2.0	18.75	36.5	51.75	77.0
2	Anglia Well	42.0	35.880952	21.527297	2.0	17.00	36.0	49.00	85.0
3	Appletree Islands	99.0	36.858586	21.534979	0.0	18.50	39.0	50.50	86.0
4	Baker Fortress	6.0	18.000000	17.332051	0.0	5.00	16.5	25.00	46.0
...	...	...	...	...	...	...	...	...	...
100	Wong Islands	32.0	43.406250	19.607206	9.0	27.25	42.5	56.50	82.0
101	Woods Fortress	6.0	35.833333	12.544587	25.0	28.75	33.0	35.75	60.0
102	Yates Burg	63.0	38.380952	21.933498	0.0	23.50	38.0	54.50	83.0
103	Yellow Inlet	123.0	36.918699	22.129868	0.0	21.50	34.0	51.50	89.0
104	Yucca Square	170.0	35.064706	22.168570	0.0	18.00	33.0	49.75	102.0

105 rows × 9 columns

Table 4



	House Number	Occupancy count
count	3422.000000	3422.000000
mean	45.784629	2.830801
std	52.168157	2.181144
min	1.000000	1.000000
25%	11.000000	1.000000
50%	26.000000	2.000000
75%	52.000000	4.000000
max	228.000000	22.000000

From table 3, there are 105 streets and the household density of each street. From table 4, overall household density, all of the streets have 3422 houses. There are approximately 45 houses on each street. The minimum house of each street is 1. The maximum number of the house on each street is 228. Moreover, in each house, there are almost 3 people in their family. The minimum size of the family is 1 person, and the maximum size of the family is 22 people.

Age pyramid

Figure number 2

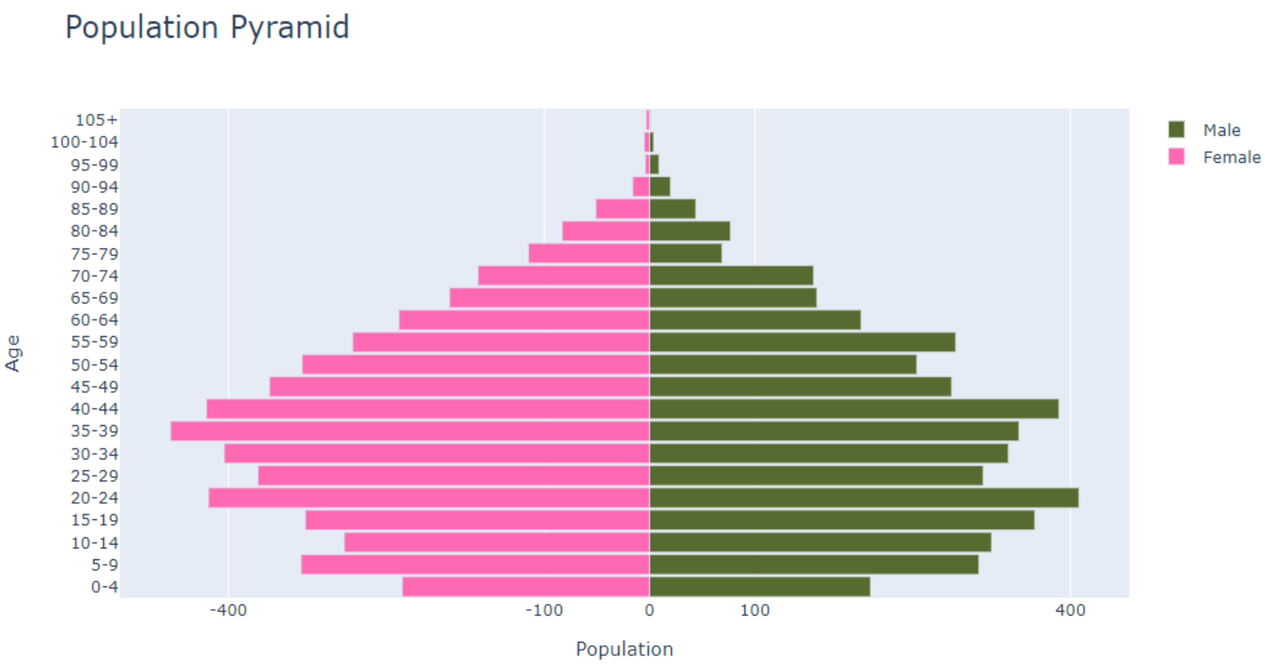
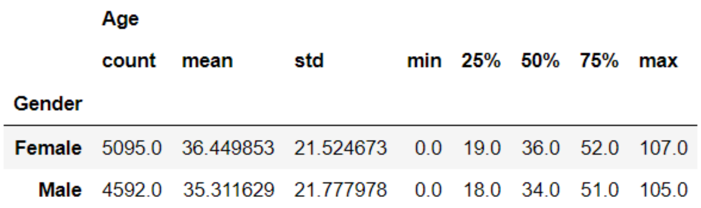
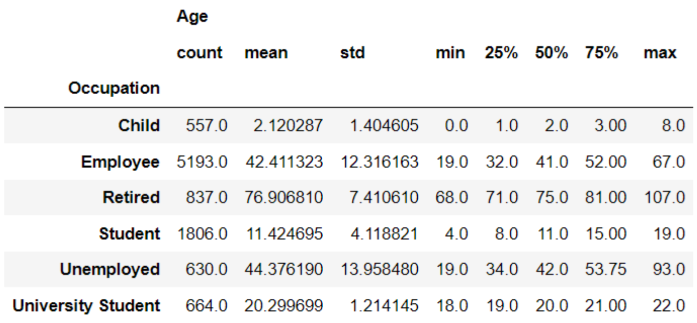


Table 5



	Age							
	count	mean	std	min	25%	50%	75%	max
Gender								
Female	5095.0	36.449853	21.524673	0.0	19.0	36.0	52.0	107.0
Male	4592.0	35.311629	21.777978	0.0	18.0	34.0	51.0	105.0

Table 6



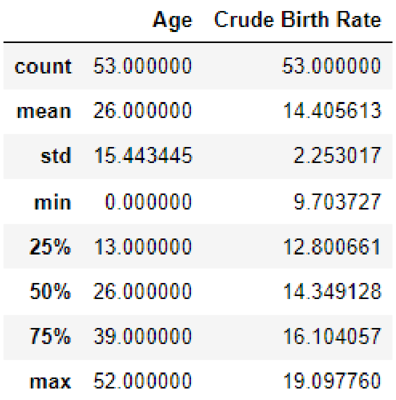
	Age							
	count	mean	std	min	25%	50%	75%	max
Occupation								
Child	557.0	2.120287	1.404605	0.0	1.0	2.0	3.00	8.0
Employee	5193.0	42.411323	12.316163	19.0	32.0	41.0	52.00	67.0
Retired	837.0	76.906810	7.410610	68.0	71.0	75.0	81.00	107.0
Student	1806.0	11.424695	4.118821	4.0	8.0	11.0	15.00	19.0
Unemployed	630.0	44.376190	13.958480	19.0	34.0	42.0	53.75	93.0
University Student	664.0	20.299699	1.214145	18.0	19.0	20.0	21.00	22.0

From table 5, this age pyramid depicts the population of the female are more than male 503 people. From figure number 2 and table 6, more and more people are children, teenagers, and adults. They may need more schools, industries, company. If someone needs to make owner from other lands to invest. They need to develop the land to be a better place and modern land. For example, public transportation such as trains should be built.

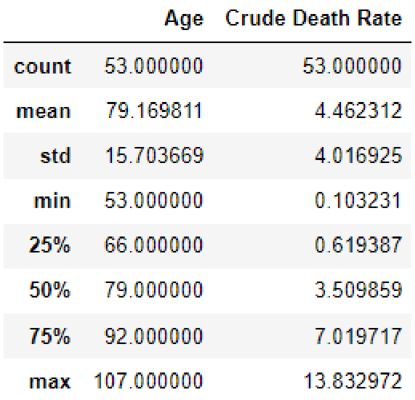
$$\frac{\#\ \text{births in 1 year}}{\#\ \text{thousand total population}} = \text{Crude Birth Rate}$$

$$\frac{\#\ \text{deaths in 1 year}}{\#\ \text{thousand total population}} = \text{Crude Death Rate}$$

Crude Birth Rate Table 7 and Crude Death Rate Table 8 (Atom, 2013)

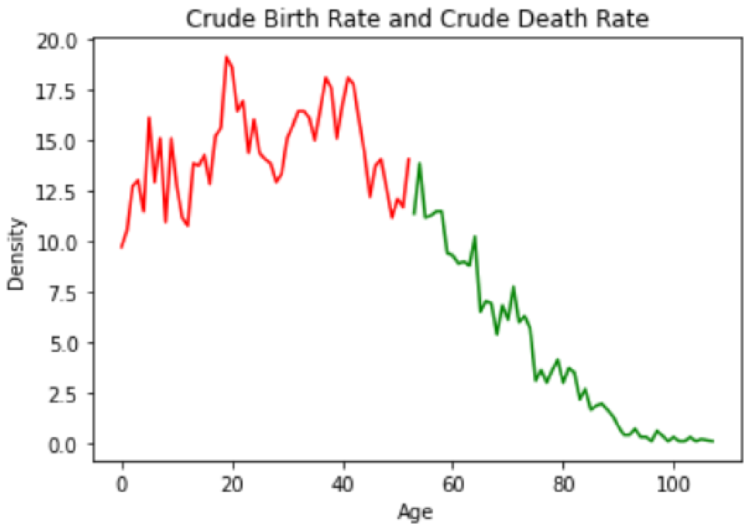


	Age	Crude Birth Rate
count	53.000000	53.000000
mean	26.000000	14.405613
std	15.443445	2.253017
min	0.000000	9.703727
25%	13.000000	12.800661
50%	26.000000	14.349128
75%	39.000000	16.104057
max	52.000000	19.097760



	Age	Crude Death Rate
count	53.000000	53.000000
mean	79.169811	4.462312
std	15.703669	4.016925
min	53.000000	0.103231
25%	66.000000	0.619387
50%	79.000000	3.509859
75%	92.000000	7.019717
max	107.000000	13.832972

Figure number 3



Both formulas show how to calculate Crude Birth Rate and Crude Death Rate. From tables 7 and 8, both data tables have been separated by half of the maximum age to 0-53 and 54-107. Both tables above show the overall number of Crude Birth Rate mean is 14 in thousand people and Crude Death Rate mean is 4 in thousand people. As a result, population density is increasing by approximately 10 in thousand people. Therefore, people's birth is more than people who pass away.

Married and Divorced lead to immigrate and emigrate Figure number 4

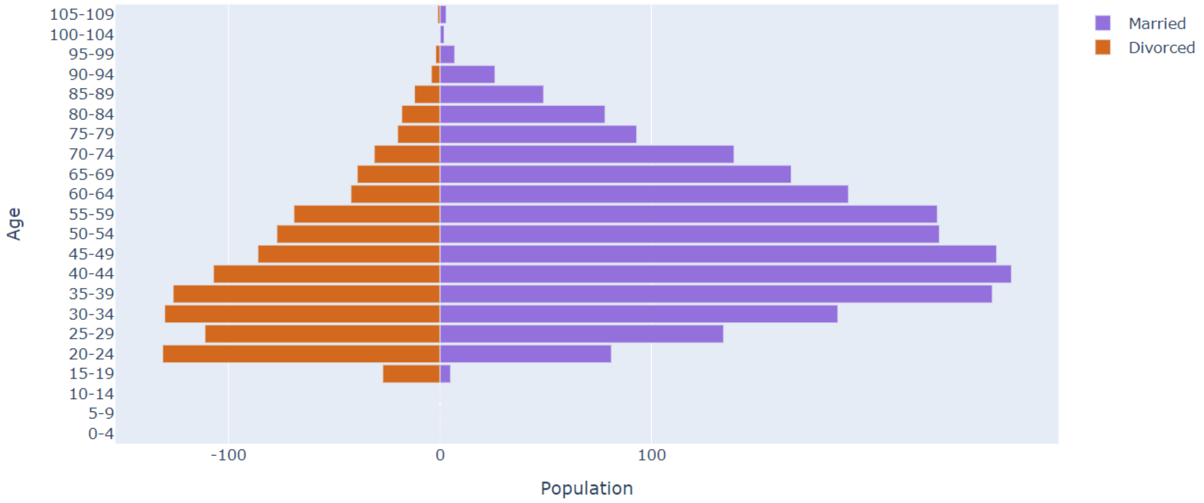
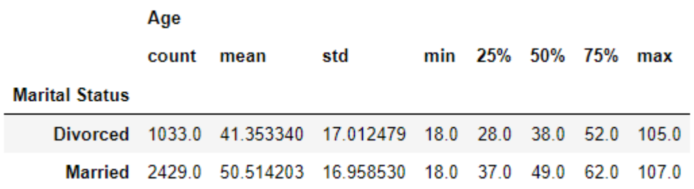


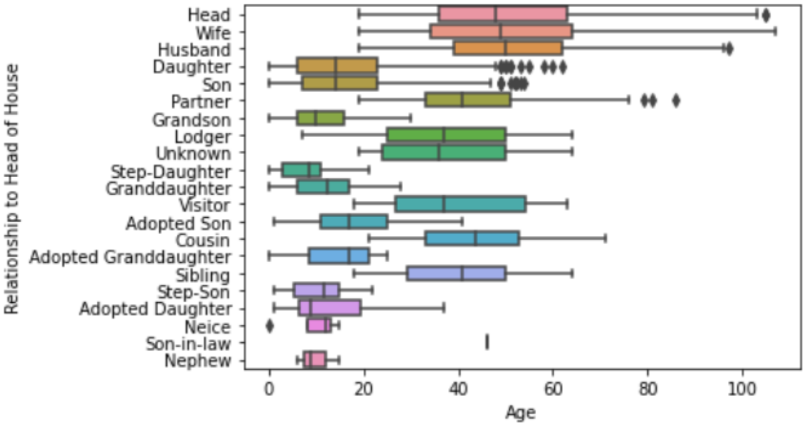
Table 9



	Age							
	count	mean	std	min	25%	50%	75%	max
Marital Status								
Divorced	1033.0	41.353340	17.012479	18.0	28.0	38.0	52.0	105.0
Married	2429.0	50.514203	16.958530	18.0	37.0	49.0	62.0	107.0

Figures number 4 and table 9 show people who get married people are more than divorced, 1396 people. Someone who is divorced will set up a new house or they are emigrating to live with their parents.

Figure number 5

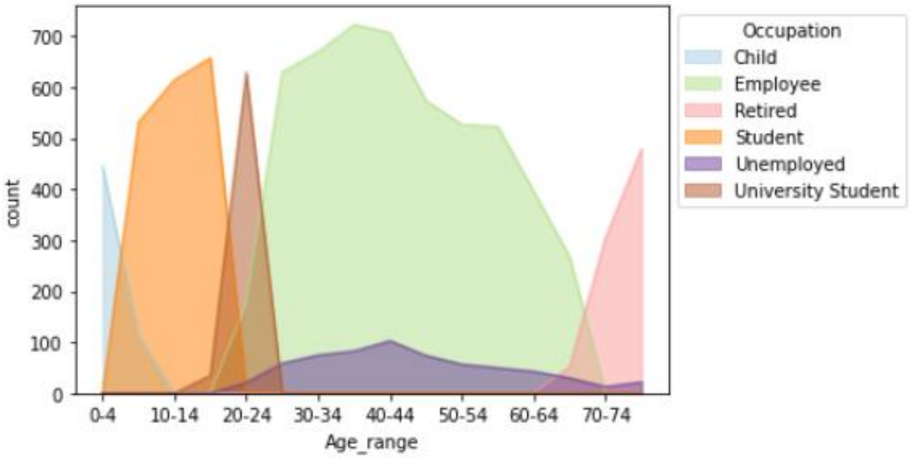


On the other hand, someone who gets married, their family will probably immigrate someone brings the whole family to this city and it can make population density is increasing. On the one hand, Someone gets married does not bring everyone to live with them, varies by culture. In comparison, from table 9, the number of married people is more than divorce. As a result, the population density will be increasing.

people's information

Employed and Unemployed

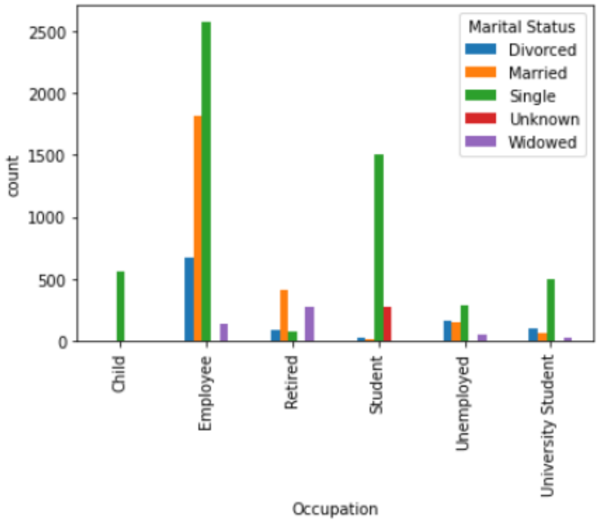
Figure number 6



From figure number 6, the relationship between occupation and age can be clearly seen in this graph. First, it shows the number of employees and unemployed are between 19 to 70 years old. Secondly, the number of employees is more than unemployed. As a result, this land has good employment and training. Moreover, many people have a good education that is currently studying in school and university.

Relationship between occupation and marital status

Figure number 7



From figure number 7, this graph shows that most people are employees, students, and university students. Moreover, they are single. Furthermore, it will be better if this land has public transportation such as a train station. Because they are single, it's more convenient to travel alone.

Religious populations

Figure number 8

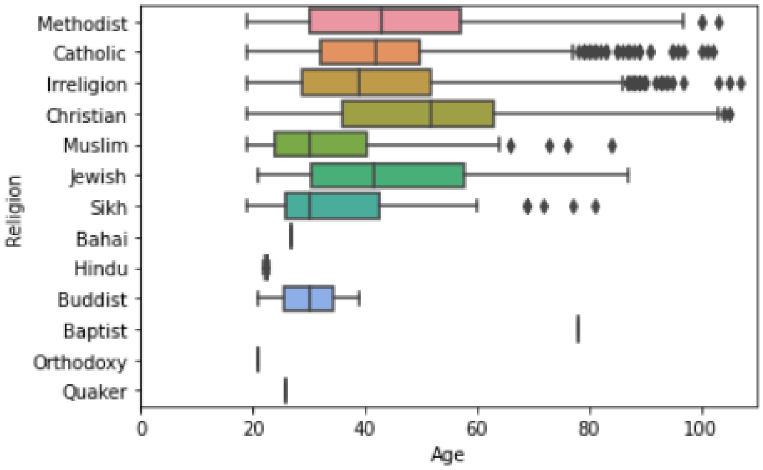
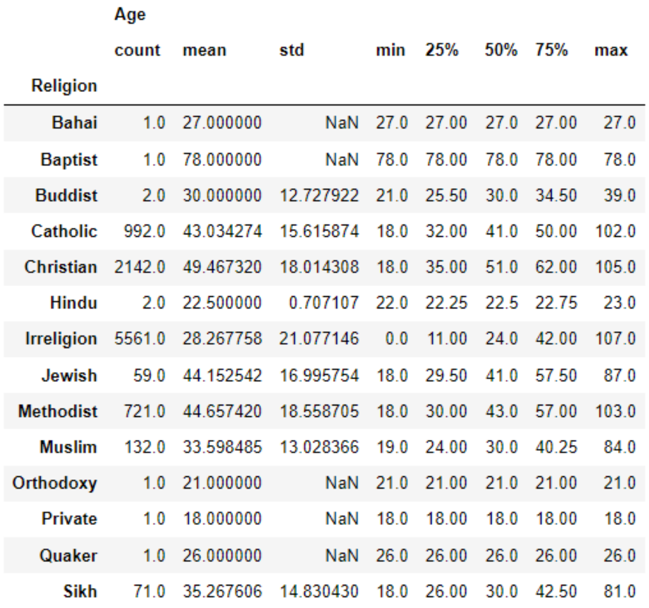


Table 10



	Age							
	count	mean	std	min	25%	50%	75%	max
Religion								
Bahai	1.0	27.000000	NaN	27.0	27.00	27.0	27.00	27.0
Baptist	1.0	78.000000	NaN	78.0	78.00	78.0	78.00	78.0
Buddist	2.0	30.000000	12.727922	21.0	25.50	30.0	34.50	39.0
Catholic	992.0	43.034274	15.615874	18.0	32.00	41.0	50.00	102.0
Christian	2142.0	49.467320	18.014308	18.0	35.00	51.0	62.00	105.0
Hindu	2.0	22.500000	0.707107	22.0	22.25	22.5	22.75	23.0
Irreligion	5561.0	28.267758	21.077146	0.0	11.00	24.0	42.00	107.0
Jewish	59.0	44.152542	16.995754	18.0	29.50	41.0	57.50	87.0
Methodist	721.0	44.657420	18.558705	18.0	30.00	43.0	57.00	103.0
Muslim	132.0	33.598485	13.028366	19.0	24.00	30.0	40.25	84.0
Orthodoxy	1.0	21.000000	NaN	21.0	21.00	21.0	21.00	21.0
Private	1.0	18.000000	NaN	18.0	18.00	18.0	18.00	18.0
Quaker	1.0	26.000000	NaN	26.0	26.00	26.0	26.00	26.0
Sikh	71.0	35.267606	14.830430	18.0	26.00	30.0	42.50	81.0

From figure number 8, this graph depicts the data who are more than 18 years old because table 10 shows people who are less than 18 years old cannot decide their religion or irreligion. In this land, more and more people are irreligion. However, there are many religions in this land. To begin with, the first one is Christianity is the highest number religion in this land, but only elder people believe in Christian. Therefore, In the future, the number of people who believe in Christian will be decreasing. On the one hand, the number of people who believe in Muslim is increasing in teenagers and adults. The land should build religious buildings such as the mosque. Last but not least, Baptist is going to shrink because no one transmits this religion from parents to children.

Factors leading to infirmity and age

Figure number 9

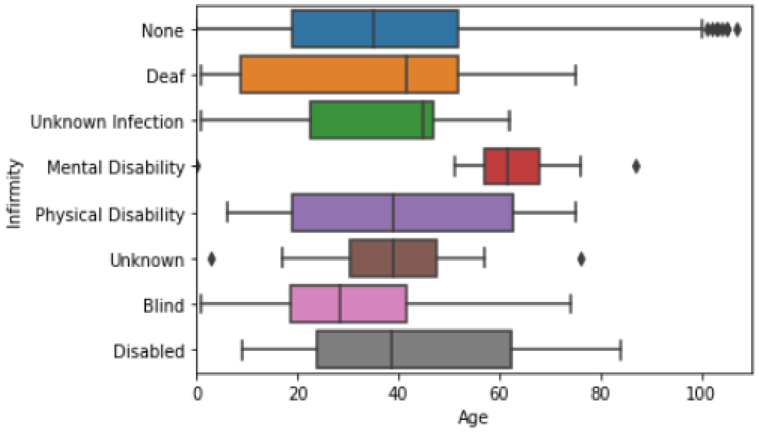
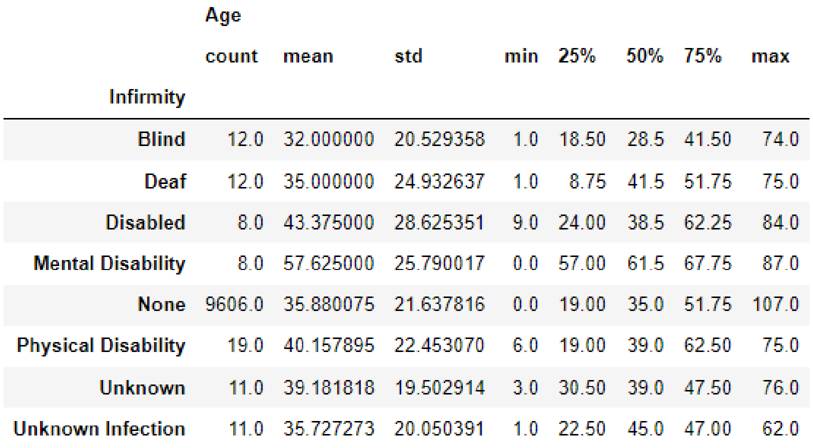
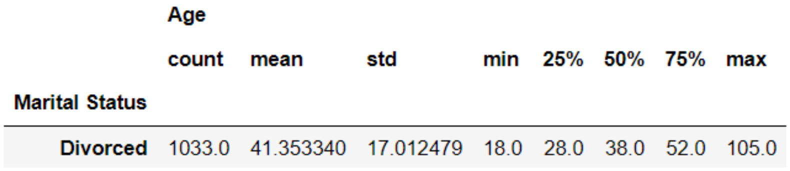


Table 11



	Age							
	count	mean	std	min	25%	50%	75%	max
Infirmity								
Blind	12.0	32.000000	20.529358	1.0	18.50	28.5	41.50	74.0
Deaf	12.0	35.000000	24.932637	1.0	8.75	41.5	51.75	75.0
Disabled	8.0	43.375000	28.625351	9.0	24.00	38.5	62.25	84.0
Mental Disability	8.0	57.625000	25.790017	0.0	57.00	61.5	67.75	87.0
None	9606.0	35.880075	21.637816	0.0	19.00	35.0	51.75	107.0
Physical Disability	19.0	40.157895	22.453070	6.0	19.00	39.0	62.50	75.0
Unknown	11.0	39.181818	19.502914	3.0	30.50	39.0	47.50	76.0
Unknown Infection	11.0	35.727273	20.050391	1.0	22.50	45.0	47.00	62.0

Table 12



	Age							
	count	mean	std	min	25%	50%	75%	max
Marital Status								
Divorced	1033.0	41.353340	17.012479	18.0	28.0	38.0	52.0	105.0

From figure number 9 and table 11, the number of people who have no infirmity is more than people who have infirmity. On the other hand, from table 12, the divorced factor can increase the number of mental disabilities (Otterstrom and Attorney, n.d.). However, the number of people who have no infirmity is more than those who have infirmity and are divorced. As a result, now it’s not necessary to invest in the medical building.

In summary

Now plan: The train station should be built because of two factors:

1. Population density is increasing.

- Crude Birth Rate is more than Crude Death Rate.

- The number of immigrating is more than emigrating.

2. people's information: More and more people in this city are employees and single. It’s more convenient compared with the bus and the cost is cheaper than private cars.

The advantages of the train station are more convenient, cheaper and friendly environment because it can carry many people per round trip.

Future plan: Moreover, more and more people in this land believe in Christian, However, teenagers and adults do not believe in Christian. They decide to believe in Muslims. In the future if the number of people who believe in Muslim is increasing. It should be built mosque.

Investigation

Now plan: Regarding the population density is increasing and this land has good employment and training. Moreover, many people have a good education that is currently studying in school and university. This is a good opportunity to invest in general infrastructure. For example, waste collection, road maintenance, and the general infrastructure lead to the land will be developed to be a better place.

Future plan: Regarding there are many middle-aged people. Moreover, there are a few people who have infirmity. The number of elders will probably increase within the next 20 years. When the time will come, it’s a good opportunity to invest in old-age care.

Bibliography

Bloom, D. (2021) *Legal age of marriage set to be raised from 16 to 18 in England and Wales*

Available online: https://www.mirror.co.uk/news/politics/legal-age-marriage-set-raised-25496345

Accessed: 05/12/2020

eFile (2021) *IRS Head of Household Filing Status*

Available online: https://www.efile.com/irs-head-of-household-tax-filing-status/

Accessed: 05/12/2020

Atom (2013) *The demographic equation*

Available online: http://www.geog100.org/p/ch-5-population.html

Accessed: 05/12/2020

Otterstrom, K. and Attorney (No date) *Mental Health Issues and Divorce*

Available online: https://www.divorcenet.com/resources/mental-health-issues-and-divorce.html

Accessed: 05/12/2020

Collins, L. (2020) *Oldest human ever documented*

Available online: https://en.wikipedia.org/wiki/Jeanne_Calment

Accessed: 06/12/2020

Colby, L. (2016) *STAR WARS: The best Jedi and Sith, ranked*

Available online: https://www.cbr.com/star-wars-best-jedi-and-sith-list/

Accessed: 06/12/2020