UNIVERSITY OF LEEDS

School of Computing
Faculty of Engineering

Assessed Coursework Header/Feedback Sheet

This form should be completed and submitted with each piece of assessed coursework.

My signature here confirms acceptance of the declaration of academic integrity printed overleaf.

Student Name:	Himanshu Jaiswal
Signature:	
SID Number:	2 0 1 5 7 7 1 6 2

You must make sure that you have read the declaration of academic integrity.

To be completed by the student submitting work
(please make sure that you complete ALL fields)

Module Code	C O M P 5 1 2 2 M
Module Title	Data Science (33862)
Coursework Number	Coursework 2
Name of Lecturer Marking Work	Professor Roy Ruddle **You must fill this in**
Deadline Date	3rd December 2021
Date Handed In	3rd December 2021

Feedback section to be completed by the marker

Section	Marks Available	Marks Awarded	Comments on sections
Total			
Late Penalty (marks to be deducted)			

General Comments on Work

[redacted]

Other members of the group!

· ADITYA PUSDEKAR

Sno: 201570275

Sign :

· ANUPAM BOSE

Sno : 201570198

Sign :

· HIMANSHU MADAN

Sno : 201501359

Sign :

Declaration of academic integrity

Further information about plagiarism, fraudulent or fabricated coursework and malpractice in University assessments is available at: http://www.leeds.ac.uk/aaandr/cpff.htm

I am aware that the University defines plagiarism as **presenting someone else's work, in whole or in part, as your own.** Work means any intellectual output, and typically includes text, data, images, sound or performance.

*(On the understanding that other members of the group have made contributions to the attached submission,)** I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

**For the submission of group work*

Penalties for late submission

University rules on penalties for late submission of coursework require 5% of the total marks available to be deducted for each calendar day that passes after the date of required submission. The deduction will be applied to the grade/mark for the coursework component concerned before any conflation with other grades/marks to give the overall result for the module. If coursework is not submitted by the end of 14 calendar days following the prescribed deadline, a grade/mark of zero should be returned for that component.

Coursework 2

Group Analysis Case Study

Introduction

This report is submitted as coursework 2 for the Data Science module taught as a part of the MSc Data Science and Analytics programme at the University Of Leeds.

In this group analysis case study,

We aim to:

- ➢ Analyse the data containing NHS Dental Statistics for England for the period **April 2018 to March 2019,** at the level of Clinical Commissioning Groups (CCGs).
- ➢ Using the data and analysis, answer the following questions:
  - ❖ What patterns are there in the number/age of patients treated?
  - ❖ What is the effect of deprivation?
  - ❖ A person aged 55 in 2019 will reach the State Pension age in 2031. What types of CCGs face the greatest shortage of dentists in 2031?

The data:

We used the following datasets for our analysis:

- **The CCG and Patient Part:**

1. nhs-dent-stat-eng-jan-jun-18-anx3-ps-prac.csv (Link to dataset)
2. nhs-dent-stat-eng-jul-dec-18-anx3-ps-prac.csv (Link To dataset)
3. nhs-dent-stat-eng-jan-jun-19-anx3-ps-prac.csv (Link to dataset)

- **The Dentist Part:**

1. NHS Dental Statistics for England 2019-20 Annex 3_Workforce csv.csv (Link to dataset)

- **The Deprivation Part:**

1. File_7_-_All_IoD2019_Scores__Ranks__Deciles_and_Population_Denominators_3.csv (Link to dataset)

The datasets are **publicly available** on the UK government’s website.

Data Characterisation:

Understanding the data:

We merged the above mentioned data files into one single dataset using the join method in python.

There are 11 columns in our final file:

1. **PSEEN_END_DATE**: Last date of the month in which the patient was seen by a dentist.
2. **CCG_CODE:** CCG code of the CCG commissioning health care.
3. **CCG_ONS_CODE:** The code of the CCG with the ONS. ONS stands for Office for National Statistics.
4. **CCG_NAME:** The name of the Clinical Commissioning Group.
5. **PATIENT_TYPE**: Whether a patient is a child or an adult.
6. **AGE_BAND:** Determines the age of the patient.
7. **PATIENTS_SEEN:** Number of patients seen in a month within a given CCG.
8. **POPULATION:** Population of the CCG.
9. **DENTIST_AGE_GROUP:** Categorisation of the dentist’s age.
10. **NUMBER_OF_DENTISTS:** number of dentists in that particular CCG.
11. **AVG_IMD_SCORE:** IMD stands for Index of Multiple Deprivation. It is to classify the relative deprivation of small areas.

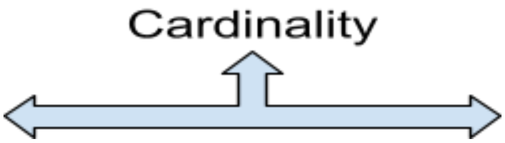
Data Preparation

Profiling the data:

Characterisation Tasks

The following table explains the variable type of each column of the data along with the minimum and maximum values by giving examples from each column. The empty cells represent that there is no minimum or maximum value for that particular column.

Columns	Data Type	Min.	Max.	Data Sample
PSEEN_END_DATE	Numeric	30/04/2018	30/06/2019	28/02/2019
CCG_CODE	Nominal	00C	99Q	99K, 07J
CCG_ONS_CODE	Nominal	E38000001	E38000230	E38000100
CCG_NAME	Nominal	----	----	NHS Darlington CCG
PATIENT_TYPE	Categorical	----	----	Adult,Child
AGE_BAND	Numeric	0	18	11,12,13..
PATIENTS_SEEN	Numeric	3	499,995	42, 398, 1022, 927
POPULATION	Numeric	0	962,934	1269, 1328
DENTIST_AGE_GROUP	Ordinal	Under 35	55+	35-44,45-54
NUMBER_OF_DENTISTS	Numeric	4	258	11,13,18..
AVG_IMD_SCORE	Numeric	7.18	52.139	9.179, 10.4



Columns	Distinct Values	Average Value	Length
PSEEN_END_DATE	12	----	10
CCG_CODE	198	----	3-11
CCG_ONS_CODE	192	----	9-11
CCG_NAME	192	----	11-57
PATIENT_TYPE	2	Child	5
AGE_BAND	19	9	1-2
PATIENTS_SEEN	8491	7903	3-8
POPULATION	2664	15164	4-9
DENTIST_AGE_GROUP	4	----	2-3+
NUMBER_OF_DENTISTS	127	43	1-4
AVG_IMD_SCORE	191	22	5-6

Data Quality:

There were several data quality issues within the data:

1. **Missing/Spurious Values:**
   1.1. Columns **CCG_CODE, CCG_ONS_CODE and CCG_NAME** have a few spurious values which can be observed as a case of missing values. Few rows are having *"UNALLOCATED"* as their value, which can be assumed to

be a case of missing values but since they are just a few, we can ignore them for our analysis.

1.2. Also, **Dentist_Age_Groups and Number Of Dentists** have missing values in some of the rows which is again a missing data issue.

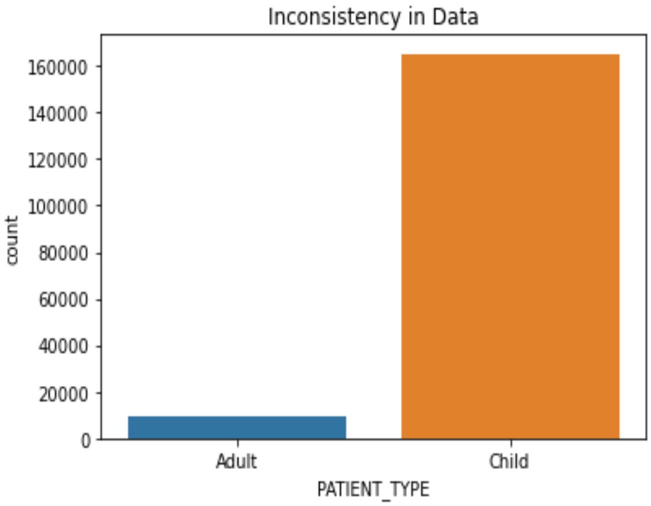
2. **Duplicate Records:**

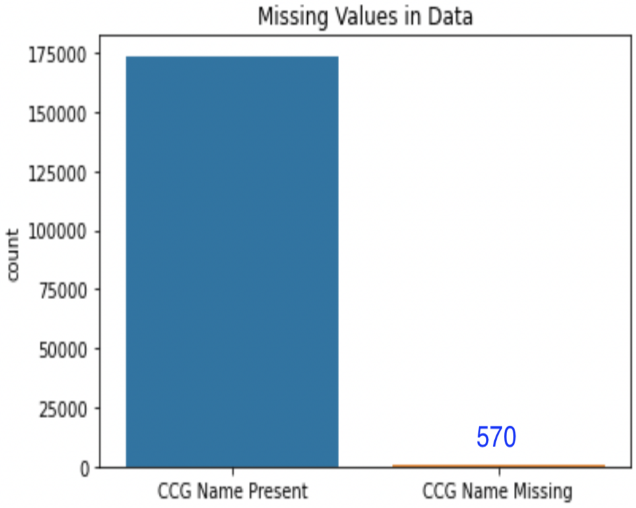
2.1. While merging the datasets, we observed that the number of Dentist's data for every CCG is not available and thus few duplicate records were formed in the dataset.

So, we performed **DATA STANDARDISATION** and cleaned the data to minimize duplicacy of records.

3. **Inconsistency:**

3.1. The number of records in the patients data shows inconsistency as for adults we have cumulative data for the past 24 months but for children, it is just for the past 12 months. So, we have to be mindful of this fact while performing analysis.





Detailed Analysis:

a). Analysis of patterns seen in number/age of patients treated :

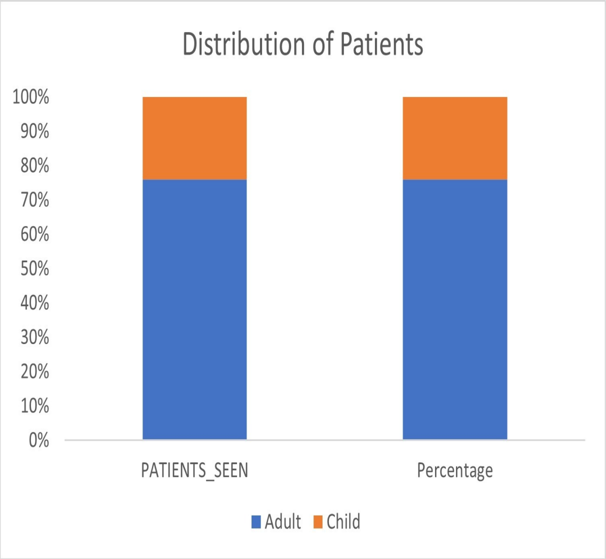


Fig. 1(a)

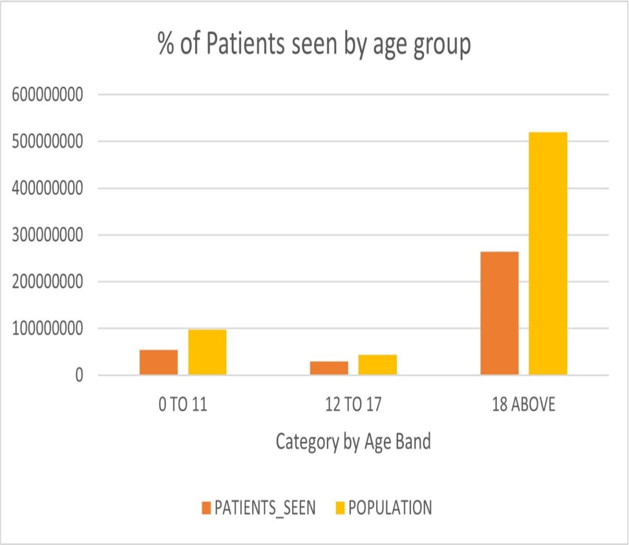


Fig. 1(b)

The bar chart Fig. 1(b) depicts the ratio of the population to the patients treated in different age bands. The population of the adults being the maximum, it makes sense from the chart that we have the maximum number of patients treated in that age group. We have the least number of patients who are children considering the small population of them. Furthermore, we could see how we have more patients in the age band of 0 to 11 than 12 to 17. It again makes sense considering the fact that babies need more care & regular health checkups in their age. Fig. 1(a) shows the distribution of patients based on the patients seen and the percentage of patients seen in adults and children. As we can see there are more adult(76%) patients than children(24%).

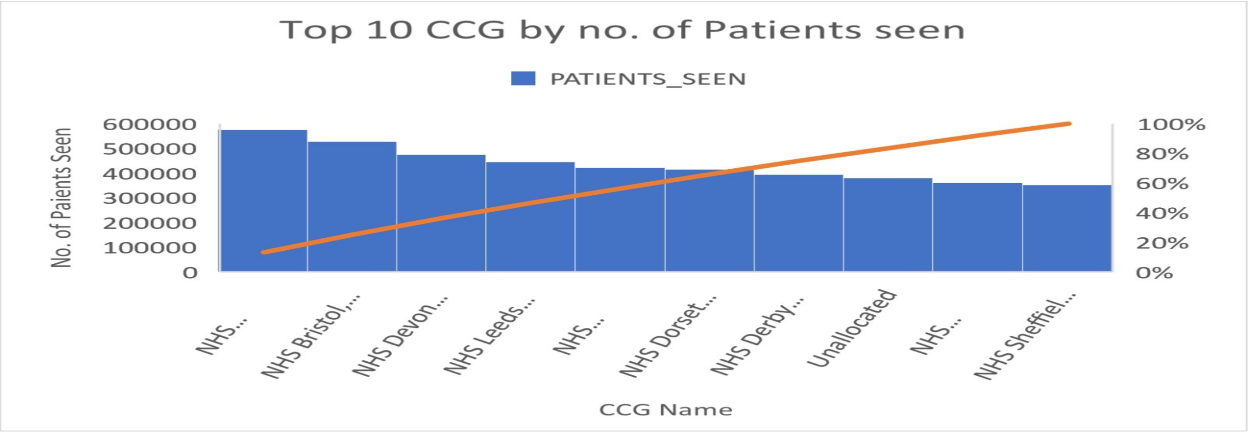
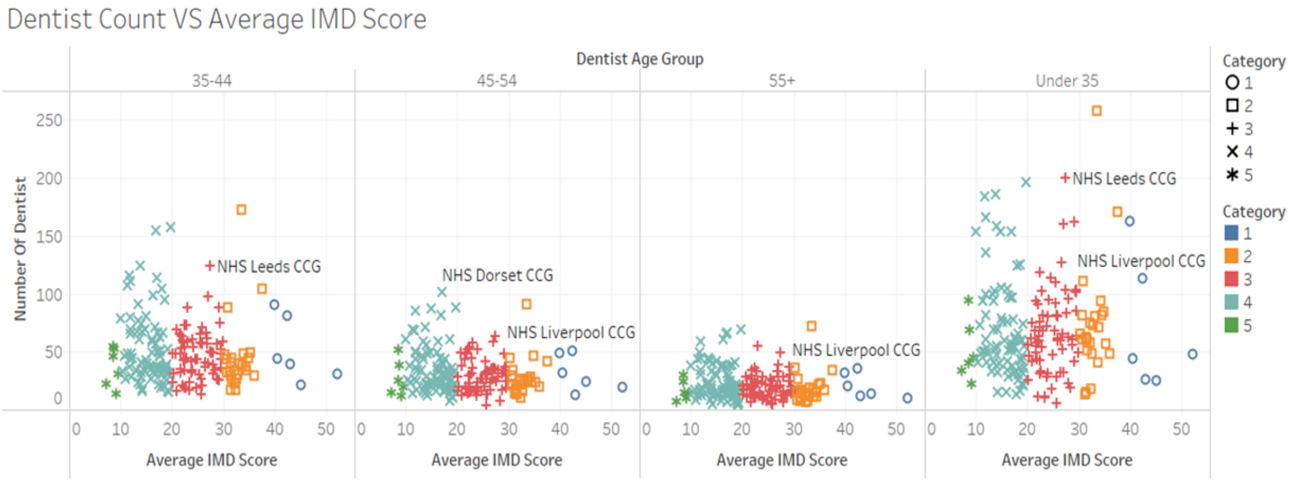


Fig. 1(c)

The graph above shows the top 10 CCGs by the number of patients seen which depicts where the most number of patients are seen in CCG's.

b). Analysis on effects of deprivation:

The average IMD Score explains how deprived the CCGs are, the score is directly proportional to deprivation in the CCGs. The following scatter plot shows different ranges of avg IMD scores of all CCGs which has been broken down into five categories. Category 1: 40 & above, Category 2: 30 to 39, Category 3: 20 to 29, Category 4: 10 to 19, Category 5: Below 10.



Sum of Avg Imd Score vs. sum of Number Of Dentist broken down by Dentist Age Group. Color shows details about Category. Shape shows details about Category. The marks are labeled by Ccg Name. Details are shown for Ccg Name, Dentist Age Group and Dentist Age Group.

Fig. 2(a)

Less the age-group of the dentist is, the more scattered the data points(CCGs) are in our scatter plot.The age group under 35 showed the highest scatter which explains how high the dentist count is with some variance & there are least number of dentists above the age of 55. We have many young dentists serving across most of the CCGs having avg. IMD score ranging from 19 to 30. Though we could analyse some of the patterns as explained above from our plot, this isn't enough to conclude that there is any dependency between the dentist count and avg IMD score and so the deprivation.

From the below graph we can infer that as the Percentage of the Patients_seen decreases the AVG_IMD_SCORE also decreases across the categories 1-5 of CCGs. Here we can also observe that as the age of the dentist increases, the dentist count decreases in each of these categories. So we can conclude that the more the CCG is deprived, the more patients have been seen in those CCGs.

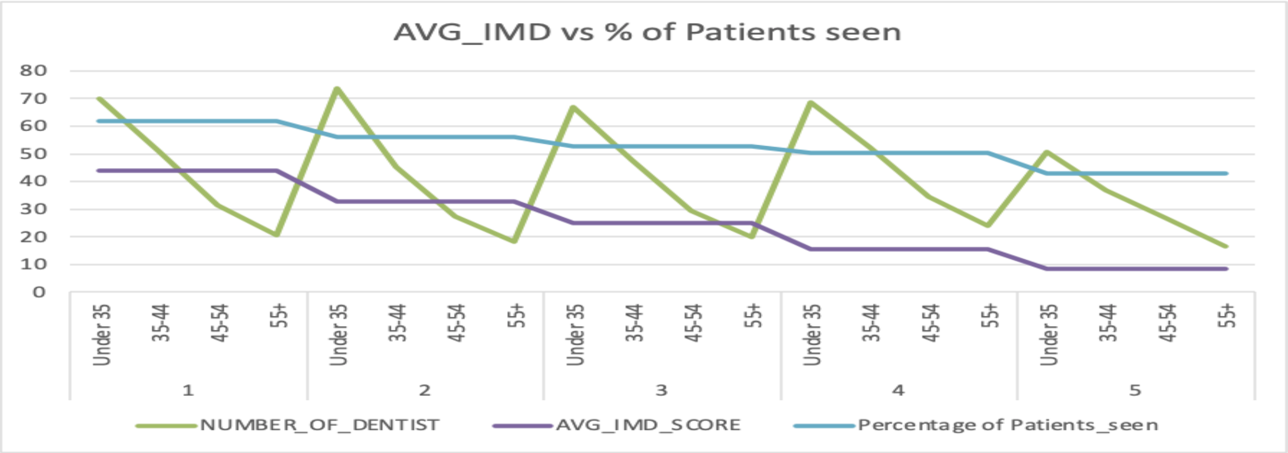


Fig. 2(b)

c). A person aged 55 in 2019 will reach the State Pension age in 2031. What types of CCGs face the greatest shortage of dentists in 2031?

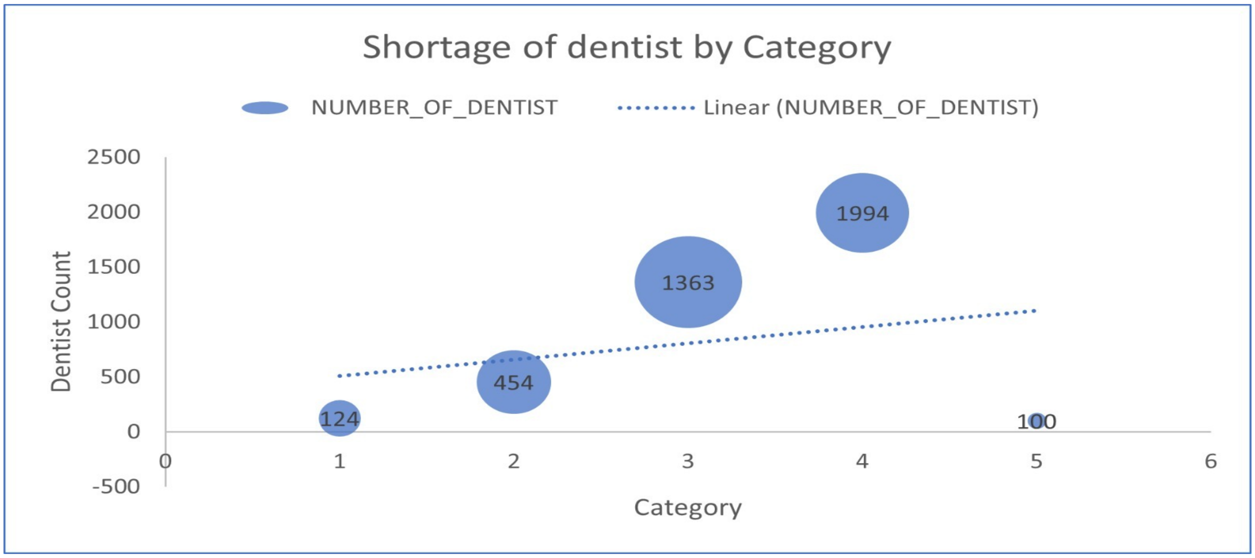


Fig. 3(a)

The graph plots the number of dentists above the age of 55 for all the 5 categories of CCGs from which we can infer that the one in which there will be the most shortage of dentists would be category 4 (avg IMD ranging from 10-19) followed by 3(avg IMD ranging from 20-29) in 2031. We can also say that the least shortage of dentists would be in category 5 (avg IMD ranging below 10) and 1 (avg IMD ranging above 40) as it is clearly evident from the graph above. In the graph below Fig. c(2) we can see that % of patients who are 55+ are mostly 30% of total dentist count which is very significant. Therefore, when these 30% of dentists will retire there will be a huge shortage of dentists in 2031.

Orange color percentage represents 55+ aged dentists and blue color percentage depicts below 55 aged dentists in 2019.

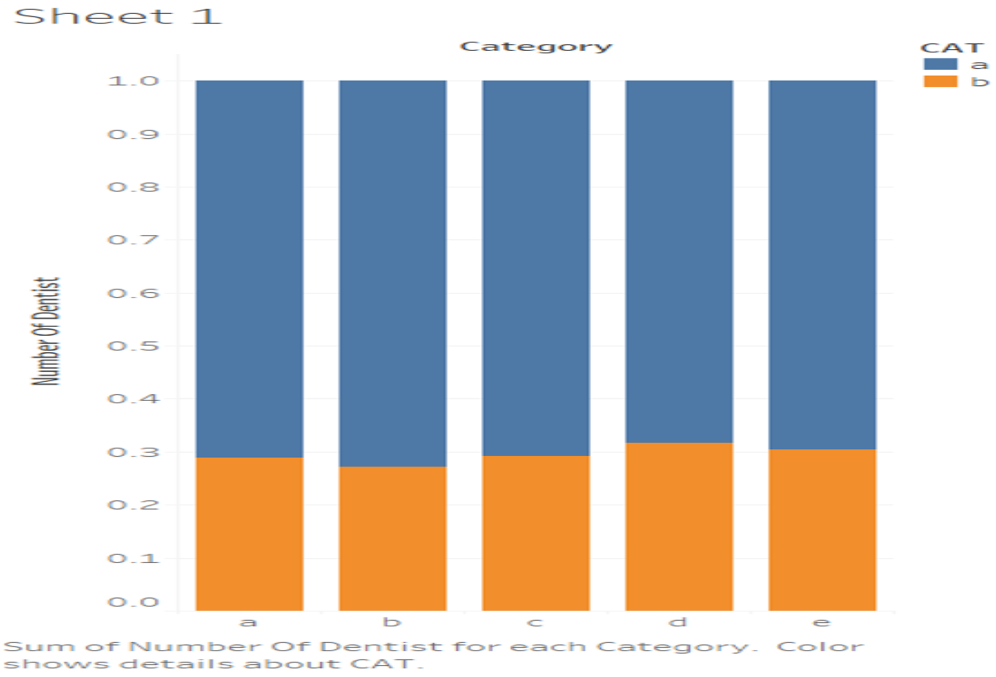


Fig. c(2)

Conclusion:

- We have inferred by performing the analysis on our data that there are maximum patients aged above 18 and least from the age group 12-17.
- We have analysed that the percentage of patients seen is directly proportional to average IMD score.
- Among the patients that have been seen, 76% are adults and 24% are children.
- We have analysed that the number of dentists are maximum for the age-group under 35 and as the age increases, the number of dentists decreases so the minimum is for the age-group above 55.
- The CCG which has average IMD score between 20 - 40 will face the most shortage of dentists in 2031.
- This data and the analysis will help NHS in future to allocate funds and dentists to ensure improvement in deprived areas, and will also be used as evidence so that the government and organization can study in which area they should provide economical and healthcare support.