**St Olave’s Grammar School**

***Department of Mathematics***

**LARGE DATA SET 5 - HOMEWORK PROJECT**

Total marks available: 43

All questions in this assignment refer to MEI Large Data Set 5, which consists of data about boroughs in London and regions of the UK. The data set is available at <https://www.ocr.org.uk/qualifications/as-and-a-level/mathematics-b-mei-h630-h640-from-2017/assessment/> under Pre-release materials.

Answer all the questions. The marks for each question are indicated in square brackets.

You are strongly advised to use computer software where appropriate.

You can submit your work in one of two ways.

* You can edit this document to include your answers. In this case, you can extend the spaces provided for your answers as you see fit, but **leave space for your teacher to add marks, comments, and annotations** to your work.
* You can print this document and write your answers by hand in the space provided.

|  |  |  |
| --- | --- | --- |
| 1. | a) Jane is asked to produce a display indicating how household recycling rates in Greater London have changed since 2004. She decides that there is too much data to display so she takes a sample of ten London boroughs to use in her work.  If she wants to be sure she represents inner and outer London accurately, what sort of sample should she take? Explain in detail exactly how she would select her sample.  She could use stratified sampling. Because the sampling frame is available, you can calculate a sampling fraction of 10/33 ≈ 0.3 . Inner boroughs make up 14 of the 33 given boroughs (including city of London), therefore 14 x 0.3 = 4.24 ≈ 4 . So Jane should sample 4 inner boroughs and 6 outer boroughs. In order to choose 4 inner and 6 outer boroughs, Jane should use a random number generator and a numbered list of each borough (inner and outer separately) so there is no bias.  b) Jane lives in Bexley and is proud of its recycling record. She says that the data shows that in 2018/19 her borough sent more than three times as much waste to recycling as the lowest performing borough.  Is Jane’s statement correct? Explain your answer.  Bexley recycles 54% of their waste, the lowest recycling rate was 17% for Newham  17 x 3 = 51 🡪 this shows the Bexley recycles more than 3x of their waste proportionally but not directly (e.g. Newham could produce much more waste while Bexley much less therefore 54% of Bexley’s waste may not be 3x greater than 17% of Newham’s in actual mass)  [part (c) is on the next page]  c) Jane produces the chart below.    Comment on her choice of graph.  Very detailed, able to show overall trend but also individual proportions of each borough for each year – therefore very good choice of graph  d) Without using a sample, but using the data from the Large Data set, produce your own graph or chart to display changes in household recycling rates in London. (You may use a computer).  Explain clearly why your display is an improvement on Jane’s.    My graph is clearer and concise in showing the trend across all the 33 boroughs (including city of London)  e) Describe briefly what your graph shows about how recycling rates have changed over the period from 2003 to 2018.  Average percent of waste recycled of all London boroughs has increased dramatically from 2003/4 -2012/13 – however this has stopped increasing and remained fairly constant from 2012/13 – 2018/19 | [4]  [2]  [1]  [3]  [1] |

|  |  |  |
| --- | --- | --- |
| 2. | a) In the data set the information about house prices is given as the median house price in each borough.  Give a possible reason why this measure of location has been chosen rather than using the mean house prices.  With median, the recorded data is not skewed/ dominated by any extreme values or outliers – therefore the median is a better representative than the mean  b) Draw a box plot to show the distribution of house prices in London Boroughs in 2018.    c) Which London Boroughs have median house prices that are considered outliers? Justify your answer with calculations.  Lower Quartile = 415000  Upper Quartile = 586250  IQR = 171250 IQR x1.5 = 256875  415000 - 256875 = 158125 -> no outliers because none are less than or equal to 158125  586250 + 256875 = 843125 -> 3 outliers of 923080, 1020000, 1425000 because > 843125  d) Draw a second box plot showing distribution of house prices in London Boroughs in 2008.    e) Compare the median house prices across London Boroughs in 2008 and 2018, referring to the box plots you have drawn.  From the boxplots, the 2018 median house price (£475000) of all the boroughs is higher than for 2008 (£265000) – almost double, showing major inflation over the 10 years. Also, the interquartile range for 2008 (£86649) is much smaller than for 2018 (171250) which means there is much more consistency in the middle 50% of the data in 2008. However, both boxplots show a proportionally similar positive skew in the data, possibly hinting towards the growing house prices. | [1]  [2]  [2]  [1]  [4] |
| 3. | a) Explain when it is appropriate to use a histogram to display data.  With numerical continuous data? – when you want to see the distribution for grouped data with different class widths  The data below is taken from the large data set.   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Mean Income of Tax Payers in 2017/18** | | | | | | | | | | | | | | | | | Mean Income (in £1000) | | 27- | 34- | 36- | | 38- | | 41- | | 46- | | 60- | | 100-180 | | | Number of boroughs | | 4 | 4 | 4 | | 4 | | 4 | | 5 | | 5 | | 3 | | |  |  | | | |  | |  | |  | |  | |  | |  | | **Mean Income of Tax Payers in 2003/04** | | | | | | | | | | | | | | | | | Mean Income (in £1000) | | 17- | 21- | 23- | | 25- | | 27- | | 30- | | 36- | | 45-90 | | | Number of boroughs | | 3 | 5 | 7 | | 5 | | 3 | | 2 | | 4 | | 5 | |   b) Draw two histograms (on separate axes) to show the data above.  2017/18    2003/04    c) Use these histograms to compare the mean income across the London Boroughs in 2003/04 with the mean income in 2017/18, commenting on location, spread and shape. You are not expected to do any further calculations.  The 2017/18 data ranges from 27 to 180 (range of 153) which is around double the spread of the 2003/04 data which was located from 17 to 90 (range of 73). However the overall shape is very similar as they both have a strong positive skew as shown by the long tail on the right for both. | [1]  [3]  [3] |
| 4. | a) Calculate the mean and standard deviation for the Employment Rate in inner London Boroughs in 2019. Calculate the same information for outer London Boroughs.  mean = sum of values/number of individuals  Inner Boroughs’ mean = 1054.9/14 = 75.35 %, Outer Boroughs’ mean ≈ 74.4 (3.s.f.) %  Inner standard deviation ≈ 5.50 (3.s.f.) % – using GeoGebra Outer standard deviation ≈ 4.14 (3.s.f.) % – using GeoGebra  b) Compare and comment on the two mean values.  The mean employment rates are very similar between the inner and outer boroughs, inner boroughs is higher by ≈ 1% - almost negligible – this shows that both inner and outer London have similar proportions of the population in employment in 2019  c) Compare and comment on the two standard deviations.  The standard deviation for both inner and outer are also very similar – difference of ≈ 1.4 – and because the standard deviation is low, this is showing a similar spread and consistency of the data for both inner and outer boroughs | [2]  [2]  [2] |
| 5. | a) Explain what is meant by “cleaning” data.  The process of detecting and correcting any incomplete, incorrect/inaccurate, or missing data  b) After first cleaning the data, draw a scatter diagram showing Median Income of Tax Payers in 2014/15 against Achievement in School in 2014/15 for the London Boroughs.    Also find the correlation coefficient between these variables.  0.374292  c) Find the correlation coefficient between Median income tax in 2014/15 and Achievement in school in 2014/15 across the different regions of England.  0.920875  d) Interpret the two correlation coefficients that you found in (b) and (c).  they both show a positive correlation however, the correlation for the regions of England is much stronger than for the boroughs of London  e) Jaipreet says that it is not surprising that there is positive correlation between achievement in school and median income, because people who do well in school go on to get higher paying jobs.  Is this a valid conclusion from the data? Justify your answer.  No, because the data for those marks in school grade achievements were in the same year as the data for the median income (2014/15) therefore is not valid to use this data to conclude something for different cohorts of the population as one (even though it is logical interpretation). | [1]  [3]  [1]  [2]  [2] |