DATA ANALYTICS SAMPLE EXAM QUESTIONS

Question 1: Short questions related to R and tidyverse

Mark each statement as **True or False**. Justify your answers and, if false, provide the correct answer.

(i) <u>The following code takes the gapminder data, and produces a scatter plot of life expectancy (lifeExp) vs GDP (gdpPercapita) where all points are coloured blue.</u> True or False? Justify your answers and, if false, provide the correct answer.

```r
ggplot(data = gapminder) +
 geom_point(mapping = aes(x = gdpPercap, y = lifeExp, colour = "blue"))
```

(ii) <u>The following dataframe is in tidy format.</u> True or False? Justify your answers and, if false, provide the code to make the dataframe tidy.

	pregnant	male	female
1	yes	NA	10
2	no	20	12

(iii) <u>The dataframe *bike* contains data on the number of bikes rented out in London. You can glimpse its structure below</u>

```
> glimpse(bike)
Observations: 3,103
Variables: 17
$ date          <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-05, 2011-01-06, 2011-01-07, 2011-01-0...
$ bikes_hired   <dbl> 4555, 6250, 7262, 13430, 13757, 9595, 9294, 9338, 10558, 16058, 16412, 13894, 15911, 14834, 11...
$ season        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ avg_temp      <dbl> 6, 3, 0, 3, 6, 4, 6, 6, 3, 3, 6, 8, 12, 10, 9, 10, 7, 3, 3, 3, 1, 3, 7, 4, 4, 4, 2, 1, 1, 2, 1...
$ avg_humidity  <dbl> 84, 79, 80, 87, 84, 92, 92, 82, 79, 87, 82, 89, 89, 87, 79, 83, 90, 85, 88, 86, 83, 86, 85, 83...
$ avg_pressure  <dbl> 1025, 1028, 1024, 1013, 1000, 996, 999, 997, 1012, 1011, 1006, 1011, 1009, 1010, 1015, 1016, 1...
$ avg_windspeed <dbl> 10, 8, 6, 6, 19, 5, 11, 23, 16, 14, 16, 16, 23, 24, 24, 23, 8, 11, 8, 8, 8, 10, 13, 11, 14, 11...
$ rainfall_mm   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ rain          <lgl> TRUE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, ...
$ fog           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE...
$ thunderstorm  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRU...
$ snow          <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE...
$ year          <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011...
$ month         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2...
$ month_name    <ord> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan,...
$ day           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26,...
$ day_of_week   <ord> Sat, Sun, Mon, Tue, Wed, Thu, Fri, Sat, Sun, Mon, Tue, Wed, Thu, Fri, Sat, Sun, Mon, Tue, Wed,...
```

<u>To create a boxplot of *bikes_hired* on a month-by-month basis, we use</u>

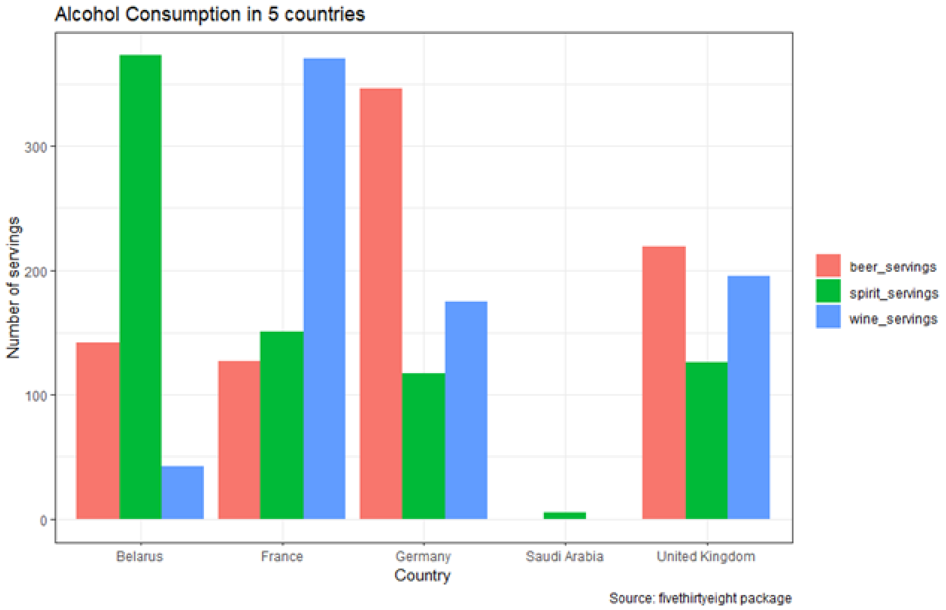
```
ggplot(data = bike, mapping = aes(x = month, y = bikes_hired)) +
  geom_boxplot()
```

True or False? Justify your answer and, if false, provide the correct answer.

(iv) <u>The *fivethirtyeight* package, has a dataset *drinks* with data on the number of servings to help us identify where people drink the most beer, wine and spirits. From this dataset se selected a few countries and the resulting dataset, *small_drinks,* is shown below</u>

	country	beer_servings	spirit_servings	wine_servings
1	Belarus	142	373	42
2	France	127	151	370
3	Germany	346	117	175
4	Saudi Arabia	0	5	0
5	United Kingdom	219	126	195

<u>Using tidyverse packages and functions, how you would you create this plot?</u>



Question 2

In determining automobile mileage ratings, we rely on the relationship between the distance travelled and the amount of fuel consumed by a vehicle that ascertains the automobiles' fuel efficiency. The measure used for this purpose is expressed in "miles-per-gallon", mpg. It was found that the mpg in the city for a certain model is normally distributed, with a mean of 22.5mpg and a standard deviation of 1.5mpg.

(i) You buy a car of this model to drive it mostly in the city. What is the probability that its mpg in the city is more than 24mpg?

(a) What is the probability that its mpg in the city is between 21.5mpg and 23 mpg?

(b) What is the probability that its mpg in the city equals to the average, i.e., exactly 22.5mpg?

(ii) Find the mileage rating that the upper 5% of the cars of this model achieve.

(iii) Suppose that the car manufacturer of this model, samples 100 cars from its assembly line and tests them for mileage ratings. What is the probability that the sample mean will be greater than 21mpg? The standard deviation of this sample of 25 cars is 1.5 mpg.

Question 3

We want to study whether there is any difference in male and female first year GPAs at US colleges and universities. We collected a sample of 1000 students and the summary statistics are given below:

	female_GPA	*male_GPA*
Mean	2.545	2.396
Standard Deviation	0.759	0.716
Count	484	516

(i) **<u>Please state what is the population, the sample, the parameter you want to infer, and the available sample statistic</u>**

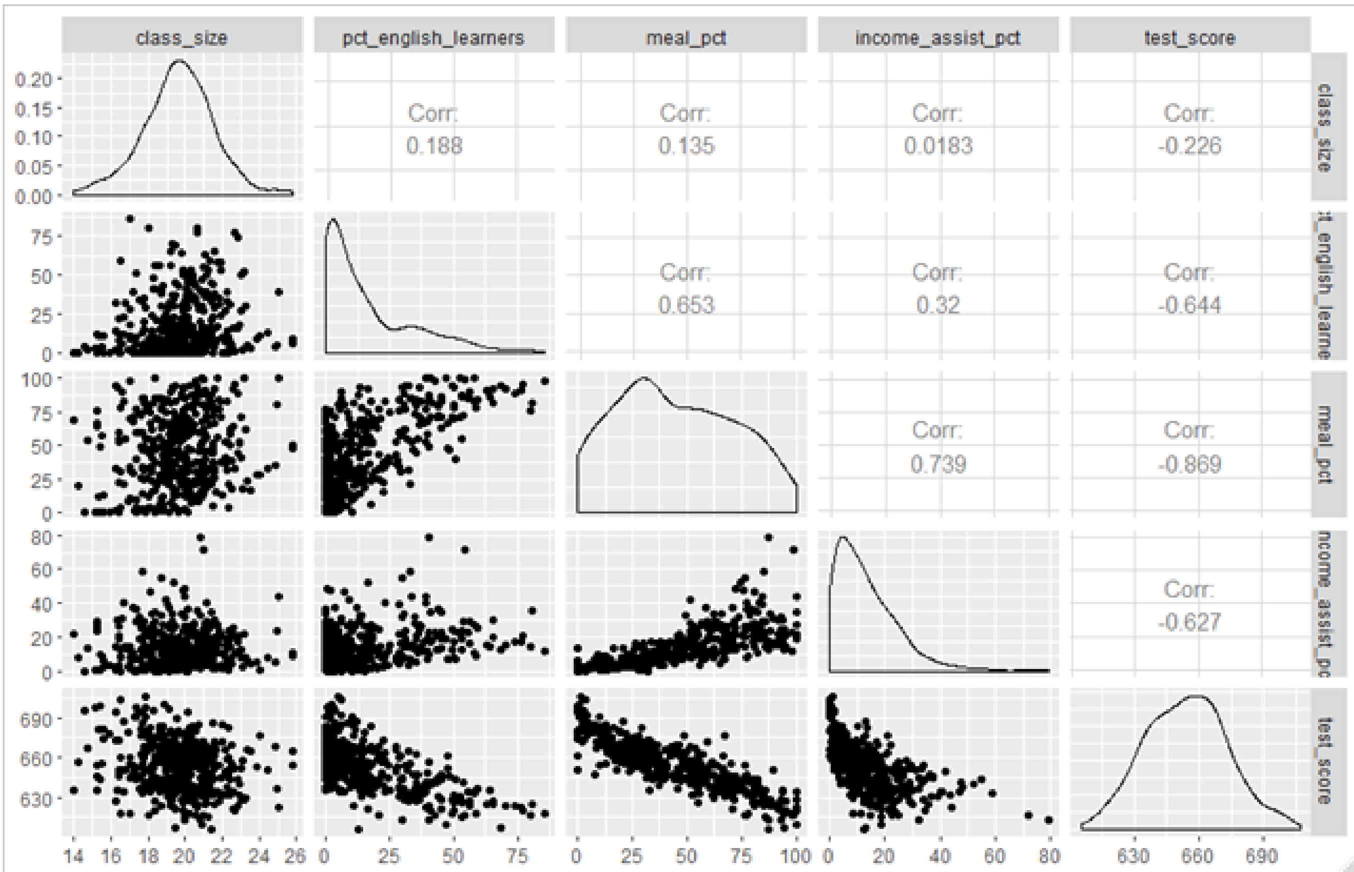
(ii) **<u>Construct two 95% confidence intervals; one for the mean female GPA and one for the mean male GPA. Do you have to make any assumptions?</u>**

(iii) **<u>Based on this sample, test whether or not the mean difference of GPAs for first year students is the same or not. Use a 5% significance level. Conduct a hypothesis testing, state the null and the alternative, calculate a t-statistic for the difference, and finally state what you decide/infer.</u>**

Question 4

We wanted to examine the effect small class sizes have on standardised test scores. We collected data from 420 elementary school districts in California and the following table shows a scatterplot- correlation matrix of all available variables

Table 3.1 Scatter plot - correlation matrix of available variables



The variables are as follows:

- **class:** average class size in school district
- **pct_enlish_learners:** percentage of students in the district for whom English is not their native language
- **meal_pct:** percentage of students in the district receiving free school meals
- **income_assist_pct:** percentage of students whose families were in an income support programme
- **test_score:** the average standardised score in the school district

In addition, in our quest to understand what explains variability in ***test_score***, we have run four regression models, 1 through 4, the summary results of which are shown in table 3.2 below.

Table 4.2 Four regression models

```
> msummary(model1)
            Estimate Std. Error
(Intercept) 698.9330     9.4675
class_size   -2.2798     0.4798

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,   Adjusted R-squared:  0.0489
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

```
> msummary(model2)
                     Estimate Std. Error
(Intercept)         686.03225    7.41131
class_size           -1.10130    0.38028
pct_english_learners -0.64978    0.03934

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264,    Adjusted R-squared:  0.4237
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

```
> msummary(model3)
                     Estimate Std. Error
(Intercept)         700.14997    4.68569
class_size           -0.99831    0.23875
pct_english_learners -0.12157    0.03232
meal_pct             -0.54735    0.02160

Residual standard error: 9.08 on 416 degrees of freedom
Multiple R-squared:  0.7745,    Adjusted R-squared:  0.7729
F-statistic: 476.3 on 3 and 416 DF,  p-value: < 2.2e-16
```

```
> msummary(model4)
                     Estimate Std. Error
(Intercept)         700.39185    4.69797
class_size           -1.01435    0.23974
pct_english_learners -0.12982    0.03400
meal_pct             -0.52862    0.03219
income_assist_pct    -0.04785    0.06097

Residual standard error: 9.084 on 415 degrees of freedom
Multiple R-squared:  0.7749,    Adjusted R-squared:  0.7727
F-statistic: 357.1 on 4 and 415 DF,  p-value: < 2.2e-16
```

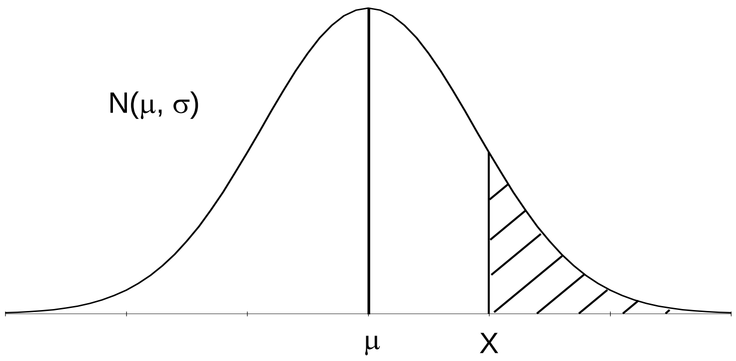
(i) **Looking at model 1, is class size a significant predictor of *test score*? What proportion of the overall variability in *test_score* does *class_size* explain**

(ii) **Consider model2. Are both explanatory variables significant? What is the proportion of variability in *test score* that is explained by model 2? What is the effect of class size and why has it changed?**

(iii) **Consider models 3 and 4. Which one do you choose and why? Given your choice, predict the test score a school district with class size = 22, pct_english_learners = 25, meal_pct = 60, and income_assist_pct = 10 is likely to get and give an approximate 95% prediction interval.**

(iv) **Looking again at your best model, the teachers' union claims that reducing class size by five (5) students, will improve test scores by at least 25 points. Do you agree or disagree with this claim?**

Table: The standard Normal distribution



Z = (X - μ)/σ	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010